Psychological Bulletin

CONTENTS

Published Bimonthly by the
American Psychological Association

Vol. 53, No. 5 September, 1956

The Psychological Bulletin contains evaluative reviews of research literature and articles on research methodology in psychology. This JOURNAL *does not publish reports of original research or original theoretical articles.*

Manuscripts should be sent to Wayne Dennis, Department of Psychology, Brooklyn College, Brooklyn 10, New York.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the "Publication Manual of the American Psychological Association" (*Psychological Bulletin*, 1952, 49 [No. 4, Part 2], 389–449). Special attention should be given to the section on the preparation of the references (pp. 432–440), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted *in duplicate.* Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail.

Reprints. Fifty free offprints are given to contributors of articles and notes. Authors of early publication articles receive no gratis offprints.

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D. C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
Menasha, Wisconsin
and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage under the provisions of Sec. 34-40 Par. (D) provided for in Section 5.38, act of February 25, 1925, authorized August 6, 1947. Printed in U.S.A.

VOL. 53, NO. 5 SEPTEMBER, 1956

Psychological Bulletin

THE WATER-JAR EINSTELLUNG TEST AS A MEASURE OF RIGIDITY

EUGENE E. LEVITT
Institute for Juvenile Research, Chicago[1]

The water-jar Einstellung test was first used experimentally by Karl Zener and Karl Duncker at the University of Berlin in the 1920's. It was formally introduced into American psychology by Luchins in 1942 (**34**). The test consists essentially of a series of simple arithmetic problems couched in terms of three water-jars each with a known maximum capacity. The S is required to manipulate the jars so as to obtain a given quantity in one of them. No other measure except the maximum capacities of the jars can be used. The first group of problems presented to S can be solved by filling the largest jar, and then emptying it twice into one of the smaller jars, and once into the other. If the jars are labeled A, B and C, the solution, which is the simplest available, follows the form $B-A-2C$. These problems are called "*set*" or Einstellung problems since their intent is to habituate the S to the $B-A-2C$ solution. An example set problem is shown below; the capacities are noted on each jar.

50	81	7	Obtain 17
A	*B*	*C*	

Immediately after the set problem, the S is asked to solve a second group of problems for which the habituated solution suffices, but which also may be solved by a more direct method, usually $A-C$, or $A+C$. These problems are called *critical*, or *test* problems. Their original purpose was to show the effect of the set engendered by the Einstellung problems. The use of the set solution for the critical problems was taken as evidence of the establishment of a set. An example critical problem is shown below.

21	51	9	Obtain 12
A	*B*	*C*	

In his 1942 monograph (**34**) Luchins added a third type of problem, the *extinction* problem. Extinction problems are amenable to solution by the direct $A-C$ and $A+C$ method, but not by the indirect $B-A-2C$ method. The extinction problem was originally conceived as an attempt to break the Einstellung; its success or failure was determined by a subsequent group of critical problems. An example extinction problem is shown below.

19	65	7	Obtain 26
A	*B*	*C*	

Luchins has consistently regarded

[1] This study was begun while the author was associated with the Iowa Child Welfare Research Station. During that time, its preparation was supported by Research Grant MH-301 from the National Institute of Mental Health, U. S. Public Health Service. The author is indebted to Dr. Milton Rokeach for advice and encouragement, and for a critical reading of the first draft of this paper. However, the author is solely responsible for all general conclusions stated in this article.

the water-jar test as a paradigm of human learning. In 1942 (**34**) and as recently as 1954 (**40**), he made clear that he believes that the primary value of the test lies in its educational implications rather than in clinical use. Nonetheless he apparently developed some kind of clinical index of rigidity from the water-jars (**37**),[2] although he has never published any norms or other developmental data. In 1948, Rokeach (**47**) published the first account of the water-jar test as an experimental rigidity measure. He reported that there was a relationship between the number of short or direct solutions of the critical problems and scores on the California E scale. The relationship was in accord with certain theoretical considerations in which the ethnocentric individual is conceived of as having a generally rigid personality structure.

In Rokeach's final form of the water-jar test, extinction problems were not used, and the set solution of the critical problem constituted the definition of rigidity. Rokeach added the *control* problem, one of Luchins' variations (**34**), in his design. A control problem is a critical which is administered *prior* to the presentation of the set problems. *S*s who failed to solve the control problem by the short method were eliminated from the experiment. The expressed purpose of this problem was to equate the experimental *S*s "in that they all demonstrate their ability to solve a critical problem by the simple method" (**47**, p. 263).

The publication of Rokeach's findings provided the impetus for a considerable amount of experimental work with the water-jar test. It also gave rise to a controversy between Rokeach and Luchins (Cf. **36**, **48**) which has not yet been resolved. A center of the dispute was Rokeach's exclusion of the extinction problem from his form of the test. Luchins has emphasized (**36**, **38**, **39**) that the extinction problem is the proper mechanism to tap rigidity since its solution *requires* a shift of method, while that of the critical does not. If rigidity is defined as "the inability to change one's set when the objective conditions demand it," only the extinction problem logically fulfills the requirements of the definition. The contention appears reasonable, especially in light of evidence (**31**) that *S*s who use the short critical solution do not work any faster than those using the indirect solution. However, the controversy cannot be satisfactorily settled by argument alone. Logic is a secondary consideration when experimental investigation is possible.

The primary purpose of the present paper is to examine the validity of the water-jar test as a rigidity measure by critically reviewing studies involving its use as such an index. It is hoped that the controversy between Luchins and Rokeach will be resolved along the way.

Since this review is concerned with the water-jar test as a rigidity measure, a number of studies (**6**, **28**, **34**, **41**, **49**, **57**, **58**) in which the test was manipulated to investigate the effect on Einstellung, but in which it was not used as a rigidity measure, will not be considered here.

Test Variations

The expression "water-jar test" has been used thus far in its generic sense, for there are actually a number

[2] Luchins' manual has been out of print since 1950, and only a limited number of copies was ever printed (Cf. **29**). The present writer has so far been unable to obtain a copy to determine whether developmental test data are presented. The test form is probably similar to those suggested by Luchins elsewhere (**35**).

of different experimental forms of the test. We can distinguish four basic types which differ among themselves with respect to the kinds of problems which make up the form. The simplest of these we have labeled the *Zener-Duncker* form;[3] it consists of a series of set problems followed by a series of criticals. The *Luchins* form consists of sets, criticals, and extinctions. A modification of this form adds a series of criticals following the extinctions. The *Rokeach* form has a control critical problem followed by sets and criticals. The *Cowen* form consists of the modified Luchins form preceded by the control critical. A modification of this form has a set problem inserted between the first criticals and the extinctions. The sequence of the various forms are shown in Table 1. A few unlisted variations have also been used occasionally.

TABLE 1

Sequence of Types of Problems in Forms of the Water-Jar Test

Form	Zener-Duncker	Luchins		Rokeach	Cowen	
	sets	sets	sets	control	control	control
	criticals	criticals	criticals	sets	sets	sets
		extinctions	extinctions	criticals	criticals	criticals
			criticals		extinctions	set
					criticals	extinctions
						criticals

In addition to the different forms of the test, there are also various operational measures of rigidity which are derived from the forms.

Most of the water-jar studies use one or more of four primary measures. A fifth measure, time of solution of a problem, can be applied to any of the four, though in practice it is usually an extinction. These measures are listed below in the form in which a high score indicates rigidity. The individual experimenter may compute the opposite or non-rigidity score. Each measure is followed parenthetically by an appropriate abbreviation which will be used to represent it in the remainder of this paper.

1. Number of critical problems solved by the long, or indirect method (*Cr*).

2. Number of failures to solve extinction problems (*Ex*).

3. Number of long critical solutions and number of extinction failures pooled into a single score (*CrEx*).

4. A specified number of long critical solutions and extinction failures used as a multiple cutoff point to distinguish a "rigid" group of *S*s (*Cr*+*Ex*).

Considering combinations of form and measure, nine different operational definitions of rigidity (exclusive of type of administration) based on the water-jar test have been used in correlational studies. Two of these definitions involve time measures. Since the number is small, these will be included with their respective form-measure combination in compiling results, leaving only seven definitions.

The volume of rigidity studies is too limited to furnish a conclusive evaluation of the predictive validi-

[3] The forms are identified by the name of the person who first used each one, as nearly as can be determined from the literature.

ties of so large a number of definitions. However, there are sufficient studies to permit at least an inspectional comparison. To facilitate this comparison, thirty-one investigations of the relationship between the water-jar test and criterion indices have been classified according to significance of results. A study is classified as *positive* if more than 75 per cent of the reported correlations were significant at the .05 level or beyond. If less than 25 per cent were significant, it is classified as *negative*. The remaining studies are considered *ambiguous*.

A breakdown of the significance of results as a function of form, measure, and method of administration, each independently, is shown in Table 2. A similar breakdown according to combinations of form and measure is given in Table 3. Studies making up the frequency in each category of Table 3, and in the form analysis of Table 2, are listed parenthetically alongside the frequency.

The frequencies in both tables are obviously too small to warrant a statistical analysis, but inspection suggests that no particular form, measure, type of administration,[4] or combination of form and measure is superior to any of its fellows. This cautious conclusion derives additional support from the fact that two of the three positive experiments using the Rokeach form (**53**, **54**) stem

[4] Experience with the WJT suggests that it is quite sensitive to various conditions of administration, especially the instructions to *Ss*. However, these conditions are not specified in many of the studies, so that evaluation of their effects is not worth attempting.

TABLE 2

Breakdown of Results According to Form.* Measure and Administration

Form	Frequency of			Total
	Positive Results	Ambiguous Results	Negative Results	
Zener-Duncker	0	3 (8, 21, 31)	3 (20, 43, 59)	6
Luchins	1 (50)	5 (1, 5, 17, 42, 45)	7 (2, 4, 23, 26, 30, 32, 39)	13
Rokeach	3 (47, 53, 54)	1 (15)	3 (22, 27, 55)	7
Cowen	1 (16)	1 (25)	3 (14, 19, 51)	5
Total	5	10	16	31
Measure				
Cr	3	5	9†	17
Ex	1	1	4†	6
CrEx	2	2	4	8
Cr+Ex	1	0	3	4
Total	7	8	20	35
Administration				
Group	1	6	9	16
Individual	1	2	5	8
Not Stated	3	2	2	7
Total	5	10	16	31

* The bibliographic numbers of the studies in each category of the form-breakdown are shown parenthetically.
† One is a time measure.

TABLE 3

Breakdown of Results According to Combinations of Form and Measure[*]

Combination	Frequency of			Total
	Positive Results	Ambiguous Results	Negative Results	
L (*Cr*)	0	0	3 (**1**, **23**[†])	3
L (*CrEx*)	1 (**42**)	2 (**17**, **42**)	2 (**4**, **26**)	5
L (*Ex*)	1 (**1**)	1 (**5**)	4 (**23**, **30**[‡], **32**, **42**)	6
L (*Cr*+*Ex*)	1 (**50**)	0	2 (**2**, **39**)	3
Total	3	3	11	17
C (*Cr*)	0	1(**25**)	0	1
C (*CrEx*)	1 (**16**)	0	2 (**14**, **19**)	3
C (*Cr*+*Ex*)	0	0	1(**51**)	1
Total	1	1	3	5
Z-D (*Cr*)	0	3 (**8**, **21**, **31**)	3 (**20**, **43**[‡], **59**)	6
R (*Cr*)	3 (**47**, **53**, **54**)	1 (**15**)	3 (**22**, **27**, **55**)	7
Over-all Total	7	8	20	35

Note.—L = Luchins form. C = Cowen form. Z-D = Zener-Duncker form. R = Rokeach form.
[*] Studies providing the frequency in each category are shown parenthetically.
[†] Two separate *Cr* measures in this study; the *Cr*'s preceding an *Ex*, and those following the *Ex*.
[‡] A time measure.

from a single investigation (**52**). This fractionated dissertation also swells the positive results category for the *Cr* measure, and for "Not Stated" under type of administration.

In view of the data of Tables 2 and 3, it seems reasonable as well as expedient to regard the various operational definitions based on the water-jar tests as equivalent.[5] Test variations will hence be ignored in the discussions in the remainder of this paper, and results from different investigations will be pooled, when necessary, regardless of differences in form and measure.

Criticals vs. extinctions. The data in Tables 2 and 3 also provide a means of evaluating Luchins' contention that the *Ex* problem rather than the *Cr* is the proper unit for the measurement of rigidity. Only one of six instances of the use of *Ex* furnished a significant relationship, the lowest percentage of any of the four measures. It is represented by a single chi square based on a small group of *S*s. If we include the *Cr*+*Ex* measure[6] the ratio becomes two of ten instances, or 20 per cent compared with 18 per cent for *Cr* and 25 per cent for *CrEx*. A total of three out of twelve individual correlational analyses using *Ex* and *Cr*+*Ex* measures attained the .05 level of significance, and one of these (**50**) is a questionable result. As we shall see, this proportion is about the same as that of sig-

[5] This does not mean that the forms are equivalent in every sense. Means and variances furnished by the forms are not necessarily the same. The equivalence is one of predictiveness (or lack of predictiveness) rather than of descriptive data.

[6] Luchins (**39**) has recently stated that the *Cr*+*Ex* measure should be a satisfactory rigidity index.

nificant correlational analyses with all forms (see p. 352, below). Evidently, the extinction problem is no more predictive than the critical problem when both are evaluated against independent criteria.

The experiments of Guetzkow (**23**) are usually cited as experimental evidence that the *Cr* and *Ex* are different measures. He reported that men and women *S*s perform similarly on the *Cr*'s, but that a significantly higher proportion of men solved the *Ex*. On this basis, Guetzkow suggests that there are two different processes and causal factors involved; the *Cr* is concerned with acquisition of set, while the *Ex* is involved with surmounting the set. The data themselves offer no real basis for such a notion. They merely show the differential performance of the sexes. Seventy-eight per cent of the *S*s who used the short *Cr* solution also solved the *Ex*. Guetzkow feels that this lends weight to his different process idea since there was no sex difference in the 78 per cent. This is indeed a peculiar conception. If a fair proportion of the sample manifests both types of behavior, then there is likely to be a substantial correlation between solutions to the two kinds of problems. A correlation of .53 between the two problems has been reported elsewhere (**32**). Such results can hardly lead one to conclude that there is "a clear and verified distinction" between the processes involved in solving the two types of problem. Furthermore, median test analyses of Harris' data (**24**) show that when time of solution is the measure, there are no differences between males and females on either problem. In any event, the data of the correlational analyses indicate that the *Ex* is no better than the *Cr* as a unit for the measurement of rigidity, no matter what other hypotheses one might care to entertain.

The Water-Jar Test as a Rigidity Measure

In addition to suggesting the equivalence of the sundry definitions based on the water-jar test, the data of Tables 2 and 3 also indicate that the water-jar test (WJT) lacks predictive validity. Sixteen of the thirty-one studies report negative results. An additional ten have ambiguous findings. Only five can be considered as supporting the claims for the test as a rigidity measure, and it must be remembered that two of these are offshoots of a single dissertation (**52**).

The number of individual correlational analyses varies considerably among the 31 studies. Eleven studies have only a single correlation, but some report as many as 30 or more. There is a total of 202 individual analyses of the relationships between the WJT and other tests. These involve 111 measures derived from 66 different instruments, exclusive of tests of intelligence. Of the 202 correlations, 151, or 74.75 per cent do not reach significance at the .05 level or beyond. If we allow for the ten correlations which are probably significant by chance alone, the percentage of insignificant correlations rises to about 80, roughly the same as the percentage of negative and ambiguous studies shown in Table 2.

The high proportion of insignificant correlations, like the similar proportion of negative and ambiguous studies, indicates that the WJT lacks criterion validity. It is pertinent to inquire, however, whether any particular criterion test has been found to be more consistently related to the WJT than others. Evaluation is hindered by the fact that only 12 of the 66 criterion tests have been investigated in replicated studies. Nonetheless, it may be revealing to examine the relationship of the WJT and individual criteria.

The California scales. Of the 12 tests in replicated researches, the California E and F scales have been the objects of most attention, probably as a joint result of the use of the E scale in Rokeach's provocative study (**47**) and the popular theory linking rigidity and the antidemocratic personality. In view of the usually high correlation between the two scales, they will be considered as one measure for the analysis of this section.

There are nine studies (**4**, **8**, **17**, **20**, **27**, **31**, **32**, **47**, **59**) in which either the E or the F scale has been used as a criterion measure. A total of 1,088 *S*s have been involved. The results of these investigations are summarized in Table 4.

There are 18 individual correlational analyses in the studies.[7] Fifteen of these are correlation coefficients of various sorts while three are *t* tests of differences between mean WJT scores for ethnocentric and nonethnocentric groups, or between E- or F-scale means for rigid and nonrigid groups on the WJT. Of the 18 correlational analyses, only five reach the .05 level of significance or beyond. Assuming that the various coefficients are equivalent estimates,

TABLE 4

REPORTED RELATIONSHIPS BETWEEN THE WJT AND THE CALIFORNIA SCALES OF THE ANTIDEMOCRATIC PERSONALITY†

Study	*N*	Correlational Technique	Special Experimental Conditions	Result
4	29	tau	week after stress	−.15
	26		day after stress	.34*
	24		none	−.21
8	80	r	none	.00
	82		ego-involvement	.40**
17	33	r	none	.06
20	50	r_{bis}	none	.06
	50		stress	−.10
27	50	r_{bis}	none	−.23
31	20	t	none	+significant
	37		reward incentive	nonsignificant
	16‡	rho	none	−.13
	23‡		reward incentive	.09
32	29	tau	none	−.18
47	70	t	none	+significant
59	262	r	mild stress	−.03
	135		none	.07
	72§		none	.30**
Average of correlation coefficients				.04
Average of all analyses (see text)				.07
Average of all analyses of data obtained under special conditions				.05

* Significant at the .05 level.
** Significant at the .01 level.
† The signs of some correlations have been changed so that a positive correlation always indicates that WJT rigidity scores and authoritarianism scores vary in the same direction.
‡ Only *S*s with WJT scores greater than zero included.
§ A group of female Naval officers.

[7] Results for a group of children in Rokeach's study (**47**) are not included here in the interests of uniformity of research populations.

the average of the 15 is .04. If we assume further that an insignificant *t* equals a coefficient of zero, and that a significant *t* equals a coefficient of .40, the average correlation becomes .07. Evidently the WJT and the California scale of the antidemocratic personality are not related indices.

The data in Table 4 also provide a means of testing Brown's (**8**) hypothesis that a significant relationship between authoritarianism or ethnocentrism and the WJT is a function of stressful or ego-involving experimental conditions. There are seven correlations (**4**, **8**, **20**, **31**, **59**) computed from the data of 509 *Ss* who performed in this type of atmosphere. Only two of these are significant, one being furnished by Brown himself. The average using the crude conversion from *t* to *r* as in the previous paragraph is only .05. The average for the remaining 11 correlations is .08. (The respective averages without the converted *t*s are .06 and .02.) The data thus do not support Brown's hypothesis, even when his own results are included among them.

Further analysis of the results indicates that stressful experimental conditions also do not increase the predictiveness of the WJT for measures other than the *E* and *F* scales. Of 28 such correlational analyses only one is significant at the .05 level. Apparently, the data will not bear out a conclusion that the WJT's lack of criterion validity is a function of experimental circumstances.

The Rorschach. Four studies have been concerned with the relationship between the WJT and various Rorschach indices. A total of 25 such indices have been used in the studies; however, only 12 will be found in replicated studies. Of these, 11 (high *F*%, low *R*, high *A*%, high *F*+%, low *FC*, high *W*, high *Dd*, low *M*, low *FY*, low content range, and slow reaction time for first response to each card) comprise the Fisher Rigidity Score. In the Applezweig (**4**) and Katz (**27**) reports, the Fisher score is given as a unit. Both report nonsignificant relationships with the WJT. Cowen and Thompson (**15**) used 8 of Fisher's 11 measures with child *Ss*, and found that 4—*R*, content range, time of first response, and *F*+% were related to the WJT.

Both Katz and Cowen and Thompson found a significant relationship between the WJT and *M*+*C*. Of 8 indices, this is the only one found to be related by Katz. Those which he

TABLE 5

Subject Loss Due to Experimental Standards Reported in Sixteen Water-Jar Test Studies With an Original Total of 2,385 Subjects

Standard	Number of *Ss* Lost	Percentage of *Ss* Lost in Studies Using This Standard	Percentage of *Ss* Lost from Total Sample
Set solution of a requisite number of set problems	221	24.97	9.27
Short (or long) control solution	181	20.43	7.59
Arithmetic accuracy	119	22.54	4.99
Pooled standards*	113	24.84	4.74
Total Loss	634		26.58

* A pooled loss due to multiple standards is reported. Included in this figure are 32 *Ss* whose loss is not explained.

reports as unrelated include the Reichard Prejudice Score, the Gibby Stability Score, and judges' ratings of inflexibility and emotional construction. Cowen and Thompson used 18 different indices of which 8 correlated significantly with the WJT. However, the fact that their sample consisted of children tends to vitiate comparisons with the other studies.

Total *R* was reported to be related to the WJT by Cowen and Thompson and by Eriksen and Eisenstein (**17**). Katz, however, failed to confirm these findings.

Other measures in replicated studies. The relationship between the Wechsler-Bellevue Similarities subtest and the WJT has been investigated by Luchins (**39**) and Horwitz (**26**). The former reports a negative result while the latter gives a significant correlation of −.30.

Maltzman, Fox, and Morrisett (**42**), and French (**20**) each made two analyses of the WJT and the Taylor Anxiety Scale. Of the four analyses, only one of Maltzman's is significant.

The alphabet maze has been used in three studies. Cowen, Wiener, and Hess (**16**) report a significant *r* of .42, while the correlation in Vallance's study (**59**) is insignificant. Bakan (**5**) reports a significant *r* of .26, but it is the result of averaging four individual coefficients which are not independent, and is therefore a biased estimate. Only one of the four individual coefficients is significant.

Katz (**27**) and Schmidt, Fonda, and Wesley (**50**) have examined the Wesley Rigidity Scale in relation to the WJT. Katz found an insignificant relationship, while the latter experimenters claim to have found a relationship. However, their data analysis is questionable. They divided a group of *S*s into three subgroups on the basis of WJT scores, minimally, equivocally, and maximally rigid, and compared mean rigidity scale scores by *t* tests. Only one of the three *t* tests, that between the minimally rigid and the maximally rigid groups, was significant. This *t* test furnishes the basis for the claim that a relationship exists. In designs of this type, the proper procedure is first to compute an analysis of variance of the scores of all three groups. If a significant *F* does not result, significant individual *t* tests can not be regarded as indicating real differences. The over-all *F* for the three groups in the Schmidt *et al.* study is 2.89, which falls short of the .05 level. Therefore the *t* test upon which the relationship claim is based is specious, and the study must be regarded as having essentially negative results.

Oliver (**45**) and Horwitz (**26**) found no relationship between the WJT and mirror writing of letters and words.

Horwitz (**26**) and Eriksen and Eisenstein (**17**) related the WJT to performance on reversible figures. Both studies used the reversible staircase and the Necker cube. Horwitz also used the reversible profile. In the Eriksen-Eisenstein work, performance on the two figures was grouped into a single score. None of the correlations are significant except for the Necker cube in the Horwitz experiment. However, the coefficient of .30 is in the opposite direction from what would be expected if the WJT was measuring rigidity.

Applezweig (**4**) and Horwitz (**26**) computed correlations between the WJT and the Hidden Words Test. None of four individual correlation coefficients is significant.

Three different measures derived from Maier's two-string problem have been used as criterion indices. Adamson and Taylor (**1**) obtained one significant chi square of three in attempting to relate the WJT to a "functional fixedness" score. Guetzkow (**23**) reports two insignificant analyses, one based on a "stereotypy ratio" and the other on correctness of solution of the two-string problem.

The relationship between the WJT and level of aspiration measures on the Rotter Board has been investigated by Horwitz (**26**) and Harway (**25**). The latter employed 11 different measures, and analyzed differences in both means and variances for a rigid and nonrigid group differentiated by the WJT. Two of Harway's measures, number of unusual shifts of estimate (i.e., up after failure, down after success) and the absolute discrepancy between estimates from trial to trial, were replicated by Horwitz. His correlations are both insignificant, but Harway's *t* for the second measure reaches the .05 level of significance. Of Harway's 11 measures, four show both significant mean and variance differences. In addition to absolute discrepancy between estimates, real differences were found for the variation of estimates from the mean estimate, the average magnitude of the shifts, and the variation of the magnitude of shift from the mean. Harway also derived his 11 measures from the WJT itself and from the Hidden Words Test. There were only two significant mean differences for the WJT, and only one of the Hidden Words, a total of 7 of 33 mean differences derived from the three tests. However, there were five significant variance ratios from the WJT, and seven from the Hidden Words, a total of 16 for the three instruments. Of the seven mean differences, six also have significant variance differences. Variance differences are difficult to interpret, especially in the absence of accompanying mean differences. Certainly there is no particular theory or hypothesis concerning personality rigidity which would easily encompass variance differences. Such differences may have a real theoretical meaning, but it would be unduly optimistic to say that they reflect favorably on the validity of the WJT as a rigidity measure, especially since the same statement obviously cannot be made for the mean differences.

There are a number of tests used in WJT investigations which are not found in replicated studies, but which may be grouped under certain usual headings. Four of these are what are commonly regarded as tests of concept formation. Forster, Vinacke, and Digman (**19**) found no relationship between the Vigotsky and the WJT, and between another sorting test and the WJT. Katz (**27**) found a similar insignificance for the Wisconsin Sorting Test. Solomon's (**54**) "organization of biology concepts" scale did relate to the WJT at the .05 level.

Several investigations deal with the relationship between the WJT and emotional adjustment. Cowen and Thompson (**15**) found no relationship between the WJT and the Bell Adjustment Inventory or the California Test of Personality, though again, the use of a child population precludes comparison with other studies. Ainsworth (**2**) derived an adjustment score from a security-insecurity inventory, which turned out to be unrelated to the WJT. Meer's (**43**) results with the Maslow Security-Insecurity index were also negative. Horwitz (**26**) and Levine (**30**) compared groups of normals and psychiatric patients. In neither case were any significant differences in WJT scores obtained.

Three studies made use of abstract reasoning tests. Insignificant correlations with the WJT were reported by Forster *et al.* (**19**) for the Duncker reasoning tasks; the matchbox, cork, X ray, and "13." Solomon (**55**) found that the WJT was related to responses to only one of four science questions after a laboratory course designed to overcome the common misconceived answers. Sivers (**51**) apparently did not find a relationship between the WJT and Form A of the Abstract Reasoning Test; the data are not presented in sufficiently clear form to be certain.[5]

Two experiments dealt with instruments which may be thought of as measuring perceptual intolerance of ambiguity. Relationships between the WJT and the Mooney-Ferguson Closure Tests I and II, the Frenkel-Brunswik Changing Figures Test, and the Levy Design Preference Test were computed by French (**20**) for two groups. None of the eight coefficients is significant. Eriksen and Eisenstein (**17**) found that the WJT was related to "availability of hypotheses," i.e., the number of guesses as to the identity of objects shown tachistoscopically at subrecognition speeds.

Other perceptual tasks included the Angyl dots (**4**), Hidden Objects (**4**), and speed of recognition of tachistoscopically-presented words preceded by erroneous expectancy (**17**). Of seven correlations, only one—Hidden Objects—is significant. Two of the other correlations of the WJT with Hidden Objects are not significant.

Motor and perceptual tasks. A number of motor or perceptual-motor tasks have been used in the WJT studies. These include mirror writing, word construction, figure similarities, maze tracing, code deciphering, arithmetic speed, etc. The AAF Aviation Psychology Program Research Report No. 5 (**21**) lists seven "change of set tests" which fall into this group. Three of the seven were significantly related to the WJT, but the highest of the three coefficients was only .18. Three of five motor tasks in Oliver's battery (**45**) were related to WJT scores; the highest coefficient was .25 for the Gottschaldt figures. None of the three tasks used by Horwitz (**26**) was found to be related to the WJT. A contour-drawing test (**19**) was also unrelated.

Miscellaneous measures. Relationships reported in unreplicated studies, or in studies which do not fall into usual groupings, are of less import in evaluating the WJT. However, a number of such attempts are listed here for purposes of completeness.

Eight Thurstone scales were administered to a group of *S*s also performing on the WJT by Goodstein (**22**). None of the correlations was significant, the highest being only .03. Goodstein also found a similar absence of association for an anagrams test, and for the Shipley-Hartford Retreat Scale. Peer ratings were reported to be unrelated by Vallance (**59**). Decision time as measured by the Festinger-Wapner test did not distinguish rigids and nonrigids, either among normals or psychiatric patients (**30**). Cowen (**14**) found that the WJT failed to discriminate high and low "negative self-concept"

[5] The lack of clarity in no way reflects on Sivers' abilities. A study of the relationship between the WJT and the ART was not part of his design. The present writer has estimated the degree of relationship from data presented by Sivers for other purposes.

scorers on the Brownfain Self-Rating Inventory.

On the positive side, 20 of 33 items in Solomon's "aspects of scientific method" scale (**53**) successfully separated high and low scorers on the WJT. Brown (**8**) found that the WJT was related to n Achievement under ego-involving conditions, but not in an ordinary experimental situation. Solomon (**56**) reported that stutterers tended to show more WJT rigidity than nonstutterers.

Measures of intelligence. Luchins reported in his 1942 monograph (**34**) that there was no relationship between WJT scores and intelligence. No correlation coefficients or other statistical demonstrations are presented. He based his conclusion on the fact that differences in Einstellung effect varied only slightly among groups of different ages and educational levels. Such variation as did occur was attributed to "differences in attitudes towards and interpretations of their tasks and instructions, rather than sheer differences in age or educational level" (**34**, p. 19).

However, the fact that there is no correlation between mean scores of groups and intelligence does not preclude the possibility of significant correlations *within* groups. A more objective evaluation can be obtained from the results of 12 studies in which the relationships between the WJT and seven different measures of intelligence are reported. The Cowen and Thompson work (**15**) with children is again considered separately. They report no relationship with the Pintner General Abilities Test. Rokeach (**47**) found no association between either the Stanford-Binet or the Wechsler and the WJT in an adolescent group. Absence of relationship is reported by Applezweig (**4**) for the Navy GCT, and by Horwitz (**26**) for the Wechsler. Horwitz administered only three subtests, Arithmetic, Comprehension, and Similarities. The total score is unrelated to the WJT, but the latter two subtests have significant individual correlations with the WJT. Katz (**27**) found no relationship between the WJT and a composite score on Iowa Entrance Examination tests. French (**20**) found a similar lack of correlation for the AFQT.

Five studies involved the WJT and the ACE. Four of these (**5**, **7**, **43**, **45**) report significant negative relationships between the two (high rigidity, low intelligence). The remaining study (**53**) did not find a relationship. Vallance (**59**) found low but significant correlations between the WJT and academic grades in engineering and navigation obtained by students at a Navy OCS.

Again assuming that all correlation coefficients are equivalent estimates, the average correlation based on 1,218 *S*s in nine studies is −.17. Since no coefficient is reported by Benedetti and Douglas (**7**), their findings are not included. It may be reasonably assumed that the inclusion of their result would raise the average coefficient to about −.20. A small portion of the variance of WJT scores is thus probably a function of intelligence. However, this conclusion should be viewed with caution since the correlation of −.20 is largely a result of relationships obtained with the ACE, especially the "Q" sub'est. Other tests yielded mostly insignificant results, though almost all were in the right direction.

We conclude that no particular test or type of test except tests of intelligence, appear to be consistently or clearly related to the WJT. The conclusion must be tempered in light of the multiplicity of instruments used and the lack of replication. It is most particularly applicable to the California scales, which have long been considered as a rigidity criterion for performance measures. It appears to be more or less applicable to the Rorschach, to tests of concept formation, to emotional adjustment, to reasoning tests, and to various perceptual and motor tasks.

Factor-Analysis Studies

Factor analyses of batteries of tests including the WJT were performed by Horwitz (**26**) and Oliver (**45**). Apparently such an analysis was intended for the "change of set" tasks in the AAF program (**21**), but the plan was abandoned when only seven of the 28 *r*'s turned out to be significant, the largest being only .18.

Horwitz' battery included nine tests. Separate analyses were done for the normals and psychiatric patients. The results in each instance were much the same. A problem-solving rigidity factor was heavily loaded with IQ, leading Horwitz to conclude that low intelligence is "an important determinant in problem solving rigidity" (**26**, p. 70). Horwitz also derived a "strength of set" measure from the WJT by interviewing *Ss* to determine the method used in solving the water-jar problems. He grouped responses under four headings ranging from those which tended to establish the strongest set to those which led to the weakest. The "strength of set" measure was included in another factor in which arithmetic ability had a heavy loading. Hence Horwitz concludes that poor arithmetic skill accounts for the establishment of a weak set in the WJT.

Horwitz' general conclusion is that "the *Einstellung* tests appear with strong loadings on the intelligence factor but fail to cluster with any of the other rigidity tests" (**26**, p. 97). His findings are weighted further by his use of the Wechsler as an IQ measure rather than the ACE.

Oliver included ten measures in his battery, of which five were motor or perceptual-motor tasks which he had developed, three were ACE subtests, and the last was the Gottschaldt figures. Of the three factors extracted by Oliver, the WJT contributed only to General Reasoning Ability, a factor composed mostly of the ACE tests. It had a slight negative weighting for the Disposition Rigidity factor. Oliver concludes that if his Disposition Rigidity factor is validly labeled, then the WJT does not measure this characteristic. However, as in the Horwitz analysis, the WJT appears to be clearly involved with intelligence.

The factor analyses of Horwitz and Oliver support the results of the correlational studies of the WJT and intelligence tests, and lend credence to the hypothesis that scores on the WJT are in part a function of intelligence.

Performance on the WJT Under Stress

A number of studies attempt to demonstrate the validity of the WJT as a rigidity measure without the use of a criterion test. The basic design, reasoning, and intent of these studies is relatively uniform. The hypothesis under examination is that rigidity will increase as a function of stress. The *Ss* who perform under conditions of ego-involvement, anxiety, frustration, anticipated failure, and so forth, will manifest a greater frequency of long solutions to the water-jar problems than individuals to whom the test is administered under nonstressful circumstances. If the experimental results are in accordance with the hypothesis, it is customary for the experimenter to accept his results as evidence that the WJT is a valid measure of rigidity. The logic of this conclusion will be discussed in a subsequent section of this paper. For the moment, we are concerned with results obtained in experiments of this nature.

Investigations of the effects of stress on the WJT scores have been reported by Christie (**10**), Harris (**24**), Pally (**46**), and Cowen (**12**, **13**). The studies of Christie and Harris are practically replicates; both used the same design and the same WJT measure, time required to solve a single *Ex* problem following a series of sets and a *Cr*. Christie found that 15 frustrated *S*s took a mean of 157.66 seconds to complete the *Ex* while a like number of unfrustrated *S*s required only 69.87 seconds on the average. The critical ratio of the difference is reported as 2.04, with a p of .02. The p value is obviously incorrect; the t reaches only the .05 level for $df=28$. If a one-tailed test was intended, it is not so stated, nor would such a test be appropriate. In addition, the ratio of the variances for the two groups is 3.61, which is significant beyond the .01 level, and contraindicates the use of a table of Student's t for determining the p value. If we apply the adjustment for heterogenous variance suggested by Cochran and Cox (**11**), we find that the t required for significance at the .05 level is 2.14, and the difference between Christie's groups is not significant. On the other hand, the distributions of scores are skewed, a fact mentioned by Harris as leading to his use of a log transformation of the data. If we apply a median test to Christie's data, we obtain a chi square of 4.80, significant beyond the .05 level for 1 *df*.

The facts are more or less reversed for Harris' data. Using the log transformation, he finds that 18 *S*s in the stress group required 202.94 log seconds to solve the *Ex*, while the nonstress *S*s needed only 55.56 log seconds on the average. The t is given as 2.59, which is significant at the .01 level. However, analysis of the raw data using a median test results in a chi square of 2.78 which falls short of the .05 level of significance.

The interpretation of Christie's and Harris' results is thus wide open, with the particular choice depending in large part upon which statistical analysis the interpreter wishes to credit. Certainly neither study merits the unqualified citations which they have received in later publications.

Pally's study (**46**) is cumbersome, poorly reported, and difficult to evaluate. He divided his *S*s into four groups of 20 each, Groups A and B experiencing failure on tests preceding the WJT administration, Group C experiencing success, and Group D being neutral. The WJT had 10 *Cr*'s followed by an *Ex*. Once an *S* succeeded in solving a *Cr* by the short method, the experiment ceased for him. Pally proceeded to compute four measures for the groups: time required to solve the first *Cr* by the short method (if one was solved), the mean number of *Cr*'s solved, the number of *S*s having to solve the *Ex*, and the mean time required for the *Ex* solution. It is obvious that these measures are not independent of one another. There is almost certain to be a marked relationship between the number of *Cr*'s attacked by the *S* and time involved in the first short solution. Similarly, the number of *Cr*'s will be related to the number of *S*s having to do the *Ex*, and so on. Hence the significance of at least the last three measures is likely to be unclear, especially if the results are conflicting.

Pally notes that he analyzed the time for the first *Cr* scores and the number of *Cr*'s by an analysis of variance. However, no *F* ratios are reported. A table of *p* values based on *t* tests and chi squares is given, but the over-all analyses are absent. The various *p* values show that Groups A and B did not differ on any of the four measures. The same is true of Groups C and D. Group A differed from Groups C and D on three measures each. Group B differed from C and D on two measures each. In each instance, the measures providing significant differences are not the same for C and D. Both Group A and B differed from Group C on mean time for solution of the *Ex*, but neither differed significantly from Group D. However, these findings are not directly comparable to those of Christie and Harris since only 30 of Pally's 80 *S*s reached the *Ex*.

Cowen (**13**) divided his *S*s into three groups of 25 each. One group was subjected to "mild stress" prior to the WJT administration, a second group to "severe stress," while the third group was a control receiving no stress. He recorded the number of long solutions, the time of response to all problems, and the time of response to an *Ex*. In each case, the means show a clear trend from fewest long solutions, and shortest response times for the control group, to most and longest for the severe stress group, with the mild stress group in between. The three *F* ratios are highly significant. In a corollary study, Cowen (**12**) contrasted a stress group and a "praise" group. Significant differences were obtained for number of long solutions and for time to solve an *Ex*. The difference in time of solution of all problems was not significant. Cowen concludes that "less rigid behaviors were noted in the 'praise' group, presumably as a function of the anxiety-reducing effects of *E*'s praise and reassurance" (**12**, p. **427**).

This conclusion deserves some further consideration in light of the previous Cowen study (**13**). In that study, a neutral control group had a mean of 1.20 long solutions. The praise group in the second work had a mean of 3.16. The stress group of the second study (apparently the same group listed under "severe stress" in the earlier report) had a mean of 5.12. Evidently, Cowen's conclusion is *not* borne out by the data which show clearly that *both* stress *and* praise succeeded in increasing the proportion of long solutions. The same inference may be drawn from the data on time of response. The mean time for all problems for the praise group is 33.60 seconds. For the neutral control, the mean is 21.28 seconds, and only 30.36 seconds for the mild stress group! The mean time of solution for the *Ex* is 75.20 seconds for the praised *S*s, 24.72 seconds for the neutral control, and only 62.64 seconds for the mild stress group. An interpretation of this analysis is not immediately apparent. Perhaps praising an *S* for performance on a projective test and expressing interest in his further performance for correlational purposes (Cowen's "praise" technique) actually places the *S* in a stressful situation.

The study of Sivers (**51**) furnishes an interesting comparison with the five experiments discussed thus far in this section. Sivers approached the question from another angle. He distinguished a rigid and a nonrigid group on the basis of WJT scores and then selected a subsample of 44 *S*s from each group, matched on the basis of scores on Form A of the Abstract Reasoning Test. Half of the *S*s in each group were subjected to stress, after which all *S*s took Form B of the ART. An analysis of variance of the difference scores between the two forms showed a highly significant variance due to stress, but an insignificant variance for rigidity, and no interaction. In other words, stress interfered with performance on Form B, but the effects were uncomplicated by rigidity. The abstraction ability of the rigid *S*s was no more impaired by stress than that of the nonrigid *S*s. The implications of the Sivers study relative to the findings of Christie, Harris, Pally, and Cowen, will be discussed in the next section.

The investigations of Brown (**8**), Applezweig (**4**) and French (**20**), though primarily correlational studies, also furnish comparisons of descriptive data under stress and nonstress conditions. Differences between mean WJT scores under stress and nonstress are not significant in the Brown and French experiments. Appelzweig reports mean scores for three small groups of *S*s to whom the WJT was administered at different times. One group took the test a day after experiencing stress, another a week after, and the third without stress at all. The week-after group had the smallest number of short solutions and the unstressed group had the largest. The critical ratio of this difference is given as 2.04 with a p beyond the .05 level.

Applezweig also reports a significant variance ratio for the scores of the two groups. As in the case of Christie's data, the ordinary table of probability cannot be used. The Cochran-Cox adjustment raises the *CR* required for the .05 level of significance to 2.06, so that Applezweig's *CR* actually falls short. Furthermore, as is often the case with WJT data, Applezweig's distributions are markedly skewed (Cf. **3**). If we apply a median test to the data of the unstressed and week-after-stress groups, a chi square of only 0.84, $p=.26$, is obtained.

If the results of the Brown, French, and Applezweig studies are averaged, the over-all mean score for the stress group is 2.75 short solutions, and 3.00 short solutions for the nonstress group. This difference is not likely to be significant. Certainly it is far smaller than those to be found in the Cowen studies. We may reasonably conclude that the descriptive data of the correlational studies do not seem to support the conclusion that stress is accompanied by an increase in long WJT solutions.

Stress and the Validity of the WJT as a Measure of Rigidity

Some of the experimenters who have used the WJT in stress studies carefully phrase their results in terms of "problem-solving rigidity" or some similar expression. The inference, whether expressed more or less overtly or allowed to remain implicit, is that rigidity is a function of the situation rather than of the personality. Nonetheless, in discussion sections subsequent to experimental results, these same experimenters will make generalizations from situational to personality rigidity. Problem-solving rigidity is viewed as a "paradigm of maladaptive behavior," (**13**, p. 518), or it is regarded as "the same as that observed clinically and reported in studies dealing with a variety of pathological states" (**46**, p. 352). Because of such statements, the studies of the WJT under stress were considered by the present writer to be actual efforts to demonstrate the validity of the WJT as a rigidity measure.

In summing up the various studies of performance on the WJT under stressful conditions (**4, 8, 10, 12, 13, 20, 24, 46, 51**), we could hardly say more than that the over-all picture is beclouded. The evidence certainly will not support the unqualified conclusion that there is a greater degree of WJT "rigidity" manifested under stress than under nonstress circumstances. But let us assume, for purposes of discussion, that this conclusion is warranted. How does it bear on the validity of the WJT as a measure of rigidity?

To begin with, the WJT is a learning paradigm, similar in many respects to other tasks used in learning experiments. Its singular characteristic is that one of the two competing responses is made dominant, but the weaker response is the "correct" one. Studies of the effects of stress on this type of learning have been carried out by Castaneda and Palermo (**9**), Farber and Spence (**18**), and Montague (**44**) among others. The findings are summarized in the following quotation:

> If the habit strength of the correct response should be relatively weak, an increase in drive should further increase the strength of the incorrect tendencies relative to the correct tendency, resulting in impaired performance. Furthermore, the degree of impairment should be a positive function of the number and strength of the competing incorrect response tendencies (**18**, p. 120).

A translation of the Hullian dialect into water-jar terms leads to this statement: "The administration of set problems prior to the *Cr*'s or *Ex*'s makes the long (incorrect) solution the dominant one. When the *S* is placed in stress, the frequency of dominant responses to the *Cr*'s or *Ex*'s increases. Furthermore, it increases as a function of the number of set problems which were administered (i.e., as a function of the strength of the incorrect tendency)."

The hypothesis that stress is accompanied by an increase in long solutions is thus one which comes out of learning theory, and has been demonstrated by learning experiments. It fits the data of tasks like learning paired associates, discriminating colored lights, or pulling levers as well as it does the results with the WJT. The findings of Cowen, Christie, etc., thus have no particular bearing on the validity of the WJT as a rigidity measure *unless one is willing to accept any simple learning task of a certain type as a rigidity measure.* This inclusion would surely be unacceptable to those who regard the WJT as a personality index.

Sivers (**51**) provides admirable support for this stand. He showed that rigid and nonrigid *S*s on the WJT manifest similar impairment in performance on another task when stress is introduced. If the WJT were measuring rigidity, we would expect that the "rigid" *S*s would be more affected by stress. In other words, while the WJT functions adequately as a learning task, it fails as a diagnostic instrument.

The WJT as a Psychometric Instrument

Despite the widespread use of the WJT by psychologists, especially in doctoral dissertations, there has been only casual concern with its defects as a psychometric tool. There appear to be three such defects, any one of which would be likely to be regarded as serious by formal test constructors. Two of these—loss of subjects due to criteria for accepting an experimental protocol, and the skewness of distributions of scores—were pointed out several years ago by Levitt and Zelen (**31**). The third, unamenability of the test to estimates of reliability, has been practically ignored by experimenters. Each of these points warrants some extended discussion.

Reliability. Sivers (**51**) is one of the few experimenters who has been concerned with the reliability of the WJT. He sums up the matter concisely, thus:

> The reliability of the water jar test as a measuring instrument is difficult to establish directly. Most of the commonly used techniques do not suffice, for in the course of taking the test problem series many subjects discover that they have not always availed themselves of the direct method. Once a subject is consciously aware of what he might have done on previous problems, and if he has used the indirect method when the direct method could have been employed, he comes alert to further possibilities of this kind. For this reason, a test-retest situation is inappropriate. A split-half technique is obviously not to be considered inasmuch as test items cannot be equated (**51**, pp. 52–53).

In short, performance on a subsequent test will be likely to be affected by performance on the original test for many *S*s. Test-retest and equivalent forms are thus out of the question. Bakan (**5**), whose four *Ex*'s each required a different kind of solution, actually attempted to construct equivalent forms.[9] The correlation between forms was .42. However, Bakan's test was an atypical type, not used by any other experimenter, and it is not certain that the reliability which she obtained can be generalized to include other forms.

The assumptions necessary for the computation of statistical estimates of reliability are obviously not satisfied by a test in which the performance on any one item is apt to be affected by performance on previous items. In fact, there does not seem to be any sound way in which the reliability of the WJT could be estimated. The conscientious experimenter who makes use of the WJT must find some method of rationalizing away his inability to estimate its reliability.

Loss of subjects. One of the unusual aspects of the WJT as a psychometric tool is that its use seems invariably to lead to a greater loss of *S*s from the experimental sample than is customary in psychological research. The loss is a result of various standards of performance required by the experimenter on the preliminary problems in the test. The *S* who fails to perform in the requisite manner is eliminated from the final, crucial phase of the testing. Unfortunately, almost half of the studies do not report either the standards or the subject loss, although it is probable that the standards were applied in most, or all, cases. A few studies note the standards, but not the loss. In several instances, multiple standards were used, and only a pooled loss is recorded. Of 34 studies of adult samples, only 16 report both standards and loss, while 15 report neither.

One criterion is a *sine qua non* (in its most literal sense) of any WJT investigation—the solution of a requisite number of set problems by the long method. It is an obvious and well-demonstrated fact that performance on subsequent problems will be a function of the number of set problems solved by the long method. Hence the experimenter must necessarily see to it that all *S*s advancing to the crucial stage have solved the same, or approximately the same number of sets by the long method. Despite the evident importance of this standard, only 12 studies report its application, and

[9] The forms are not literally equivalent since the means differ significantly.

two of these do not give the consequent subject loss. In two of the remaining 10, a pooled loss due to multiple criteria is given.

In the remaining eight studies, 221 of 885 *S*s, a total of 24.97 per cent, have been lost due to this criterion. This amounts to 9.27 per cent of the total sample of 2,385 *S*s in the studies reporting both standards and loss. These data, as well as the losses due to other criteria, are shown in Table 5.

In the Rokeach and Cowen forms of the WJT, a short solution (or a long solution) of a control problem is also used as a standard. In the seven studies in which the loss due to this criterion can be assessed, 181 of 886 *S*s had to be discarded for failure on the control problem. This amounts to 20.43 per cent, or 7.59 per cent of the *S*s in all studies giving both criteria and loss.

Occasionally a criterion of "arithmetic accuracy" is applied. It is most often not clear just what the experimenter means by this expression; in some instances, the loss may be due to simple inability to add and subtract. In others, this standard is probably the same as the requirement of long solution of the sets (naturally, the long solution cannot be properly used unless the arithmetic computations are accurate). One-hundred and nineteen *S*s or 22.54 per cent of 528 *S*s were lost in two studies due to this criterion. This is 4.99 per cent of the over-all sample.

In four studies where the loss due to multiple criteria is given as one figure, or where there is an unexplained loss, a total of 113 *S*s of 455 —24.84 per cent—were lost.[10] This amounts to 4.74 per cent of the whole sample.

Over all, 634 *S*s, or 26.58 per cent of the 2,385 *S*s who were originally tested were eliminated from the final phase of the experiments. The percentage tends to be much higher when younger *S*s are tested. One hundred and sixty-three of 286 child *S*s in the studies of Rokeach (**47**) and Cowen and Thompson (**15**) had to be eliminated, a loss of 57 per cent of the sample.

Nor can the loss be attributed to group administration of the test. In 10 studies using such administration, 25.39 per cent of the *S*s were lost, while in four reports of individual test administration, 32.57 per cent were lost. The loss was 22.66 per cent in two studies which did not note the type of administration.

The over-all loss of over 25 per cent in the adult studies is a sizable attrition, and might very well result in a sampling bias. And losing one out of every four *S*s halfway through an experiment is hardly economical of time and research populations.

The distribution of WJT scores. A test constructor ordinarily strives to develop an instrument which will provide a normal, or nearly normal distribution of scores. His aim may be linked to theoretical considerations, but more importantly, a normal distribution enables the experimenter to apply parametric statistics —the most powerful available—in analyzing results. Regardless of the test, the individual experimenter rarely obtains true normality since most samples are relatively small. However, if the curve of the distribution is symmetrical about a maxi-

[10] An indeterminate number of these 113 *S*s were lost due to absences from testing sessions, or failure to volunteer to continue with the experiment. These losses, of course, cannot be attributed to the properties of the WJT. In all probability, the number of *S*s thus lost is quite small, and a correction for the loss would change the over-all loss only slightly.

mum ordinate at the mean, or even if it is asymmetrical but not markedly skewed, the experimenter will usually assume normality in the parent population so that he can resort to a parametric analysis. (The exact nature of the population distribution is seldom known.) But when the obtained distribution is multimodal, J shaped, or plainly skewed in some fashion, the assumption of underlying normality becomes untenable, and parametric statistics are inappropriate. An instrument which regularly leads to skewed distributions of scores in individual experiments is hence undesirable experimentally.

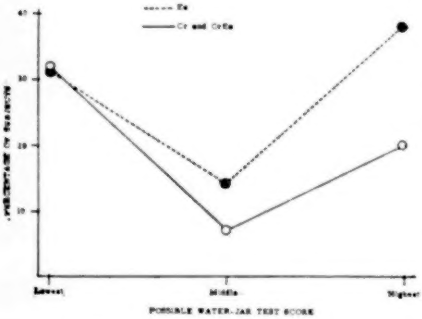

FIG. 1. DISTRIBUTIONS OF WATER-JAR TEST SCORES SHOWING PERCENTAGES OF *S*s OBTAINING THE HIGHEST, LOWEST, AND MIDDLE SCORE OF THE RANGE OF POSSIBLE SCORES. THE *Cr*-*CrEx* CURVE IS BASED ON 442 *S*s IN SIX STUDIES. THE *Ex* CURVE IS BASED ON 166 *S*s IN TWO STUDIES.

In an analysis of the data of four studies, Levitt and Zelen (**31**) called attention to the fact that the WJT furnishes an inordinately large number of zero scores. They suggested that distributions involved were probably badly skewed. In the studies involved in the present review, six others (**8**, **16**, **20**, **21**, **24**, **30**) specifically mention obtaining skewed distributions. Brown (**8**) and French (**20**) note that variance data are not presented because of skewness. Maltzman *et al.* (**42**) and Applezweig (**4**) appear to be cognizant of maldistribution in their data by the use of tau, a nonparametric correlation coefficient. Harris (**24**) reported that he converted his time measures into logarithms since the distribution of raw data was skewed.

Nine studies (**2**, **3**, **10**, **24**, **25**, **27**, **31**, **32**, **39**) either present the distributions of scores, or give sufficient information from which the shape of the distribution may be inferred. Of these, the distributions of time measures in the works of Harris (**24**) and Christie (**10**) are both clearly nonnormal. The range of possible scores in the studies using *Cr*, *CrEx* or *Ex* measures varies from 0–4 to 0–7, so that direct comparability is not simply accomplished. A reasonably clear picture of the nature of combined distributions may be obtained by plotting the percentages of *S*s attaining the highest possible score, the lowest possible, and the score at the midpoint of the range. Figure 1 shows such a composite curve for the *Cr* and *CrEx* data from six researches (**2**, **3**, **25**, **27**, **31**, **39**), and a similar curve of *Ex* scores from two studies (**32**, **39**).[11]

In the former group, 31.8 per cent of 442 *S*s received the lowest possible score, which is zero in all instances. An additional 20.6 per cent attained the highest possible score, while only 7 per cent fall at the midpoints. More than 50 per cent of all the *S*s manifest "all-or-nothing" rigidity. The *Ex* curve is similar with 31.3 per cent of 166 *S*s having a zero score, 37.3

[11] In some instances, the experimenter reported the combined frequencies for the two extreme scores at either end of the range, but did not separate them. The present writer divided the frequency evenly between the two scores in those cases. In view of the over-all data, this procedure probably tends to minimize the frequency of *S*s at the extremes.

per cent attaining the highest possible score, and 13.9 per cent at the midpoints. Here, more than two-thirds of the *S*s fall at the extremes. The difference between the percentages falling at the extremes in the *Cr-CrEx* studies and the *Ex* studies is probably a function of the more limited ranges of possible scores in the latter group. At best, we must conclude that fully half of all scores in WJT studies are likely to be found at the ends of the distribution of possible scores. Distributions will thus tend to be markedly U shaped, definitely nonnormal.

To summarize this section, the available evidence indicates clearly that the WJT is deficient as a psychometric tool in three important respects: (*a*) no reliability coefficient can be estimated for it, (*b*) about 25 per cent of *S*s originally sampled must be discarded along the way, and (*c*) it leads to skewed distributions of scores, precluding the use of parametric statistical analyses.

Methodological Shortcomings of the WJT Studies

The WJT seems to be so poor an instrument that an incautious experimenter will be easily led to commit errors of design and analysis. Researches with the WJT are rife with such flaws, some of which have been mentioned in previous sections of this paper. The most common of these is the use of parametric analyses, which are usually inappropriate, as the discussion of the last section shows. Statistics like the *t* test, *r*, and biserial *r* were used in 23 studies.[12] Other errors include the use of *S*s who failed to solve the specified number of set problems in the crucial phase, and failure to increase small *N*'s to compensate for losses when individual test administration was used.

There were a number of other shortcomings which are not directly attributable to the test itself. Among these were the use of chi square with nonindependent frequencies or with very small theoretical frequencies, failure to correct small chi-square frequencies for continuity, inappropriate use of one-tailed tests of significance, failure to use the over-all *F* test when more than two groups were involved, incorrectly stated probability levels, failure to adjust the probability level when variances were heterogenous, and inadequate explanations of the arrangement of data for statistical analysis.

That the evidence reviewed here fails to demonstrate the validity of the WJT as a rigidity measure appears to be an unchallengeable conclusion. However, many of the studies are methodologically poor, so that it is possible to argue that the WJT has not yet been subjected to sound investigation, and that any conclusion should hence be held in abeyance. The adoption or rejection of this stand is left to the reader's discretion.

Summary and Conclusions

Thirty-one correlational studies involving the water-jar Einstellung test and criterion measures were reviewed. Although there are various forms of the test, and various measures derived from it, no one was predictively superior to the others. Studies using the extinction problem as a measure of rigidity obtained no better results than those using the critical problem, or combinations of problems. Only five studies of the 31 report positive results. About 75

[12] Use of the *F* ratio with nonnormal distributions is not always inappropriate in view of evidence (**33**) that the distribution of *F* is insensitive to the shape of the parent distributions of variates involved.

per cent of over 200 individual correlations are not significant; the average of 18 correlations between the WJT and the California E and F scales is .07. Brown's hypothesis that the relationship between authoritarianism and rigidity is a function of stress or ego-involving conditions is not borne out by the data. The average of seven correlations computed from data obtained under stress is only .05.

Analysis of the relationships between the WJT and the Rorschach, measures of emotional adjustment, concept formation, reasoning, and perceptual and motor tasks indicates that no individual index has a clear or consistent association with the WJT. An analysis of nine studies of intelligence and the WJT lead to the tentative conclusion that there is a consistent, low negative relationship between the WJT and intelligence. This conclusion is supported by two factor-analysis studies, both of which place the WJT in factors heavily loaded with intelligence, though not in factors termed rigidity.

Five noncorrelational studies of the WJT in experimental stress conditions are reviewed and criticized. The results of at least three must be regarded as ambiguous, while those of Cowen (**12**, **13**) indicate that both stress and praise may increase "rigidity" on the WJT. The correlational studies of the WJT and stress in which comparative descriptive data are presented, do not support the conclusion that WJT "rigidity" increases under stress. The study of Sivers (**51**) suggests that though test performance may be impaired by stress, the impairment is unrelated to scores on the WJT. A theoretical explanation of the effect of stress on WJT scores is offered. This explanation derives from learning experiments and learning theory, and attempts to show that increases in WJT "rigidity" under stress have no bearing on the validity of the WJT as a rigidity index.

Three deficiencies of the WJT as a psychometric instrument are discussed. It is concluded that, (*a*) a reliability coefficient cannot be estimated for the test, (*b*) about one of every four *S*s in an original sample will be eliminated from the crucial experimental phases due to various standards of performance required in preliminary stages, and (*c*) the WJT tends to produce nonnormal, usually U shaped distributions of scores. A number of methodological defects in the studies reviewed were also pointed out.

The conclusions of this review can be summarized pithily in two statements:

1. After eight years of research, evidence for the validity of the water-jar test as a measure of validity is still lacking.
2. The water-jar test is a poor psychological test *qua* test.

BIBLIOGRAPHY

1. Adamson, R. E., & Taylor, D. W. Functional fixedness as related to elapsed time and set. *J. exp. Psychol.*, 1954, **47**, 122–126.
2. Ainsworth, L. H. Rigidity as a manifestation of insecurity. Unpublished master's thesis, Univer. of Toronto, 1950.
3. Applezweig, Dorothy G. An investigation of the interrelationships of several measures of rigidity under varying conditions of security. Unpublished doctor's dissertation, Univer. of Michigan, 1951.
4. Applezweig, Dee G. Some determinants of behavioral rigidity. *J. abnorm. soc. Psychol.*, 1954, **49**, 224–228.
5. Bakan, Rita. An analysis of two instru-

ments used to measure rigidity in problem solving. Unpublished master's thesis, Michigan State Univer., 1955.

6. Benedetti, D. T. The influence of freedom of choice upon mechanization in problem-solving. Unpublished doctor's dissertation, Univer. of Colorado, 1952.

7. Benedetti, D. T., & Douglass, H. O. An exploratory study of rigidity in problem-solving. Paper read at Rocky Mt. Branch of APA, Albuquerque, N. M., 1953.

8. Brown, R. W. A determinant of the relationship between rigidity and authoritarianism. *J. abnorm. soc. Psychol.*, 1953, **48**, 469–476.

9. Castaneda, A., & Palermo, D. S. Psychomotor performance as a function of amount of training and stress. *J. exp. Psychol.*, 1955, **50**, 175–179.

10. Christie, J. R. The effects of frustration on rigidity in problem solution. Unpublished doctor's dissertation, Univer. of California, 1949.

11. Cochran, W. G., & Cox, Gertrude M. *Experimental designs.* New York: Wiley, 1950.

12. Cowen, E. L. Stress reduction and problem-solving rigidity. *J. consult. Psychol.*, 1952, **16**, 425–428.

13. Cowen, E. L. The influence of varying degrees of psychological stress on problem-solving rigidity. *J. abnorm. soc. Psychol.*, 1952, **47**, 512–519.

14. Cowen, E. L. The "negative self concept" as a personality measure. *J. consult. Psychol.*, 1954, **18**, 138–142.

15. Cowen, E. L., & Thompson, G. G. Problem-solving rigidity and personality structure. *J. abnorm. soc. Psychol.*, 1951, **46**, 165–176.

16. Cowen, E. L., Wiener, M., & Hess, Judith. Generalization of problem-solving rigidity. *J. consult. Psychol.*, 1953, **17**, 100–103.

17. Eriksen, C. W., & Eisenstein, D. Personality rigidity and the Rorschach. *J. Pers.*, 1953, **21**, 386–391.

18. Farber, I. E., & Spence, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45**, 120–125.

19. Forster, Nora C., Vinacke, W. E., & Digman, J. M. Flexibility and rigidity in a variety of problem situations. *J. abnorm. soc. Psychol.*, 1955, **50**, 211–216.

20. French, Elizabeth G. Interrelation between some measures of rigidity under stress and nonstress conditions. *J. abnorm. soc. Psychol.*, 1955, **51**, 114–118.

21. Fruchter, J. Tests of set and attention. In J. P. Guilford & J. I. Lacey (Eds.), Printed classification tests, *USAF Aviat. Psychol. Prog. Res. Rep.* (Rep. No. 5), 1947.

22. Goodstein, L. D. Intellectual rigidity and social attitudes. *J. abnorm. soc. Psychol.*, 1953, **48**, 345–353.

23. Guetzkow, H. An analysis of the operation of set in problem-solving behavior. *J. gen. Psychol.*, 1951, **45**, 219–244.

24. Harris, R. A. The effects of stress on rigidity of mental set in problem solution. Unpublished doctor's dissertation, Harvard Univer., 1950.

25. Harway, N. I. Einstellung effect and goal-setting behavior. *J. abnorm. soc. Psychol.*, 1955, **50**, 339–342.

26. Horwitz, L. An investigation of the nature of rigidity. Unpublished doctor's dissertation, New York Univer., 1951.

27. Katz, A. A study of the relationships among several measures of rigidity. Unpublished doctor's dissertation, Univer. of Iowa, 1952.

28. Kendler, H. H., Greenberg, A., & Richman, H. The influence of massed and distributed practice on the development of mental set. *J. exp. Psychol.*, 1952, **43**, 21–25.

29. Klebanoff, S. G. Personal communication, Sept. 22, 1953.

30. Levine, D. Problem-solving rigidity and decision time. *J. abnorm. soc. Psychol.*, 1955, **50**, 343–344.

31. Levitt, E. E., & Zelen, S. L. The validity of the Einstellung test as a measure of rigidity. *J. abnorm. soc. Psychol.*, 1953, **48**, 573–580.

32. Levitt, E. E., & Zelen, S. L. An investigation of the water-jar extinction problem as a measure of rigidity. *Psychol. Rep.*, 1955, **1**, 331–334.

33. Lindquist, E. L. *Design and analysis of experiments.* New York: Houghton Mifflin, 1953.

34. Luchins, A. S. Mechanization in problem solving. *Psychol. Monogr.*, 1942, **54**, No. 6 (Whole No. 248).

35. Luchins, A. S. Proposed methods for studying degrees of rigidity in behavior. *J. Pers.*, 1947, **15**, 242–246.

36. Luchins, A. S. Rigidity and ethnocentrism: a critique. *J. Pers.*, 1949, **17**, 449–466.

37. Luchins, A. S. *Examination for flexi-*

bility-rigidity of behavior. Montrose, N. Y.: F. D. Roosevelt Veterans Hospital, 1950 (mimeographed).

38. Luchins, A. S. On recent usage of the Einstellung-effect as a test of rigidity. *J. consult. Psychol.*, 1951, **15**, 89–94.

39. Luchins, A. S. The Einstellung test of rigidity; its relation to concreteness of thinking. *J. consult. Psychol.*, 1951, **15**, 303–310.

40. Luchins, A. S. A variational approach to the role of set in problem solving. *Proc. 14th Int. Congr. Psychol.*, 1955, 215–217.

41. Luchins, A. S., & Luchins, Edith H. New experimental attempts at preventing mechanization in problem solving. *J. gen. Psychol.*, 1950, **42**, 279–297.

42. Maltzman, I., Fox, J., & Morrisett, L. Some effects of manifest anxiety on mental set. *J. exp. Psychol.*, 1953, **46**, 50–54.

43. Meer, B. The effect of ego-oriented and task-oriented instructions and personality differences upon problem solving behavior. Unpublished doctor's dissertation, Univer. of Pennsylvania, 1952.

44. Montague, E. K. Role of anxiety in serial rote learning. *J. exp. Psychol.*, 1953, **45**, 91–96.

45. Oliver, J. A., & Ferguson, G. A. A factorial study of tests of rigidity. *Canad. J. Psychol.*, 1951, **5**, 49–59.

46. Pally, S. Cognitive rigidity as a function of threat. *J. Pers.*, 1955, **23**, 346–355.

47. Rokeach, M. Generalized mental rigidity as a factor in ethnocentrism. *J. abnorm. soc. Psychol.*, 1948, **43**, 259–278.

48. Rokeach, M. Rigidity and ethnocentrism: a rejoinder. *J. Pers.*, 1949, **17**, 467–474.

49. Rokeach, M. The effect of perception time upon rigidity and concreteness of thinking. *J. exp. Psychol.*, 1950, **40**, 206–216.

50. Schmidt, H. O., Fonda, C. P., & Wesley, Elizabeth L. A note on consistency of rigidity as a personality variable. *J. consult. Psychol.*, 1954, **18**, 450.

51. Sivers, W. A. Problem-solving rigidity and cognitive efficiency under psychological stress. Unpublished doctor's dissertation, Syracuse Univer., 1953.

52. Solomon, M. D. The personality factor of rigidity as an element in the teaching of scientific method. Unpublished Ed.D. dissertation, Michigan State Univer., 1951.

53. Solomon, M. D. Studies in mental rigidity and the scientific method. I. Rigidity and abilities implied in scientific method. *Sci. Educ.*, 1952, **36**, 240–247.

54. Solomon, M. D. Studies in mental rigidity and the scientific method. II. Mental rigidity and comprehensiveness. *Sci. Educ.*, 1952, **36**, 263–269.

55. Solomon, M. D. Studies in mental rigidity and the scientific method. III. Rigidity and comprehensiveness in the normal situation. *Sci. Educ.*, 1953, **37**, 121–131.

56. Solomon, Nancy D. A comparison of rigidity of behavior manifested by a group of stutterers compared with "fluent" speakers in oral and other performances as measured by the Einstellung-effect. *Speech Monogr.*, 1952, **19**, 198. (Abstract)

57. Tresselt, Margaret E., & Leeds, D. S. The effect of concretizing the mental set experiment. *J. gen. Psychol.*, 1953 **48**, 51–55.

58. Tresselt, Margaret E., & Leeds, D. S. The Einstellung effect in immediate and delayed problem-solving, *J. gen. Psychol.*, 1953, **49**, 87–95.

59. Vallance, T. R. The rigidity-authoritarianism complex and its relation to performance in military office training schools. Paper read at East. Psychol. Ass., New York, 1954.

Received December 19, 1955.

Psychological Bulletin
Vol. 53, No. 5, 1956

THE NEURAL QUANTUM THEORY OF SENSORY DISCRIMINATION[1]

JOHN F. CORSO
Pennsylvania State University

From the time of classical psychophysics, the phi-gamma hypothesis has been widely accepted (**7, 13, 15, 18, 22, 39, 41**). The hypothesis states that, in a psychophysical experiment, the relationship of the proportion of observer responses to stimulus values is accurately described by the integral of the normal probability curve (**41**). Recently, however, this hypothesis has been directly challenged by the neural quantum[2] theory of sensory discrimination (**37, 38, 45**).

Broadly stated, the theory of the neural quantum may be considered as an attempt to explain a paradox in modern sensory psychology. How can a continuous change in environmental energy give rise to a (seemingly) continuous change in sensory experience, when it is generally agreed that sensory mechanisms are composed of discrete neural elements which follow the all-or-none law of physiology? The paradox may be resolved in one of two ways: experimental evidence must be obtained which demonstrates either that (*a*) sensory nerve action is continuous or that (*b*) the (apparent) continuum of sensory experience is discrete. The classical phi-gamma hypothesis assumes the first alternative, since psychometric functions are typically found to be smoothly sigmoidal in form. The more recent quantal hypothesis assumes the second alternative, since some psychometric functions have been obtained which are linear in form. In general terms, the latter findings have been interpreted as an indication that the change in nervous activity which leads to a discriminatory response proceeds in a stepwise manner by definite increments or quanta and that these quanta are directly reflected in the response itself.

Since the question of the best mathematical formula for representing a psychometric function, together with its underlying implications, is of central importance in the area of psychophysics, it would appear that some detailed attention should be directed toward the newer theoretical developments. The purpose of the present paper, therefore, is to present a complete account of the theory of the neural quantum of sensory discrimination and to re-examine the theory in the light of experimental evidence accumulated since its inception.

Early Notions of Sensory Quanta

In 1919, Titchener (**42**), in discussing the problems of measuring the stimulus and differential limens, stated that once the "nervous ma-

[1] This work was supported by a research grant NSF-G1285 from the National Science Foundation.

[2] The term "quantum," as used in the present paper, has a meaning entirely different from Planck's (**20**) quantum in physical theory, Hecht, Shlaer, and Pirenne's (**23**) quantum in visual theory, and Gabor's (**19**) quantum in auditory theory. In each of these instances, the quantum refers to a unit of physical energy; here it refers to a functionally distinct unit in the neural mechanisms which mediate sensory experience. Hence, "quantum" in the present sense implies a perceptual, rather than physical, unit.

chine" was started and adequate stimulation was continued, sensation would follow continuously the changes of stimulus. While not expressed in quantal terms, the basic notion underlying the discontinuities observed at the stimulus and differential limens was that every sense organ offered a certain amount of "frictional resistance" to the stimulus. Supposedly, this resistance had to be overcome before a corresponding change in the sensation would result.[3]

Boring (**8**) in 1926 attacked the problem of sensory experience more directly and pointed out that any theory based upon specific energies of nerves is a theory of sensory quanta. In hearing, according to this view, a new pitch is produced when the sound stimulus activates a different neural element. Furthermore, the sensory continuum reduces to a finite number of small steps, or quanta, corresponding to the number of discrete responsive elements in the given sense organ. This, essentially, was the problem Helmholtz (**24**) tried to solve years earlier when he compared the number of discriminable pitches with the number of rods in the organ of Corti.

In the absence of experimental data demonstrating the existence of pitch quanta, Boring (**8**) rejected the resonance theory of pitch and supported, instead, a frequency theory in which pitch depended upon the frequency of neural impulses and was, therefore, nonquantal. Loudness, however, was related to the number of nerve fibers activated by the stimulus, thereby making it quantal.

While not considering sensory quanta directly, Troland (**43**) pointed out the tendency toward "quantum theorizing" about the processes of the nervous system. It was suggested that "the all-or-none principle, as applied to nerve activity, forces us to think of the latter in terms of fixed units of influence" (**43**, p. 37).

In 1930, Békésy (**1**) presented the first experimental evidence which indicated that, with appropriate techniques, discrete sensory steps could be obtained, at least in the field of hearing. This was accomplished by presenting a standard tone of 0.3 sec. duration, followed immediately by a comparison tone of the same duration but of variable intensity. The observer reported whether or not he heard a difference between the two tones. The data of this study, plotted as percentage judged different against $\Delta I/I$, yielded rectilinear functions which were interpreted as indications of the quantal nature of differential sensitivity to intensity. Apparently, Békésy (**1**) was able to minimize sufficiently the "extrinsic variability"[4] in the experimental situation such that the "true" mechanism of sensory discrimination was finally revealed.

In 1936, Békésy (**2**) obtained further evidence on the quantal nature of sensory functions. In this case, the minimum audible pressures for pure tones were determined from about 2

[3] A somewhat similar view has been expressed more recently by Licklider (**30**, p. 1001) who contends that "in the simplest conceptual neurology, the stimulus threshold owes its existence to the effect of a small barrier . . . between successive stages in the neural processes that underlie hearing. . . . If the DF (Difference Limen) is more than a statistical artifact, the neural mechanism must function in a stepwise or quantal manner.

[4] "Extrinsic variability" refers to the variability in factors outside the specific part of the sensory nervous system critically involved in making the required discriminations in a given experiment, e.g., changes in criteria of judgment, in attention, in motivation, etc. (**35**).

cycles per second (cps) to 50 cps by alternately increasing frequency and decreasing intensity. When the audibility curve was plotted, steplike discontinuities occurred at fairly regular intervals between 4 cps and 50 cps, with the most prominent step at 18 cps.

THE THEORY OF THE NEURAL QUANTUM

The theory of the neural quantum in audition was made explicit by Stevens, Morgan, and Volkmann (**38**) and is derived from the assumption that the basic neural processes which mediate pitch and loudness discrimination operate on an all-or-none principle. These processes are assumed to involve neural structures which are divided into functionally distinct units or quanta. A further assumption of the theory states that a stimulus-increment will be discriminated whenever it excites one quantum more than the number of quanta excited by the standard stimulus at a given moment.[5]

On the basis of the assumption of the existence of neural quanta, sensory discrimination data would be expected to yield a rectilinear psychometric function. This would be accomplished theoretically in the following manner. Suppose that a certain stimulus excites completely a given number of quanta and that no stimulus energy, or residual, exists after this neural excitation has been accomplished. Then let stimulus-increments be added to the predetermined stimulus under specified experimental conditions of pitch or loudness discrimination. If the neural units are stable and constant, the increments will not excite the additional quantum required for discrimination until their magnitude reaches a certain size. Thereafter, each time that the increment is added to the standard stimulus, a just noticeable difference (j.n.d.) in pitch or loudness should occur. If the data of this theoretical model were presented in the form of a psychometric function, as shown in Fig. 1, with percentage of response plotted against incremental magnitude, 0 per cent response would be obtained up to a certain point on the stimulus scale and 100 per cent response would be obtained for all increments above this point. These expectancies, however, are not evidenced in auditory discrimination data and, consequently, an additional assumption on threshold variability must be introduced.

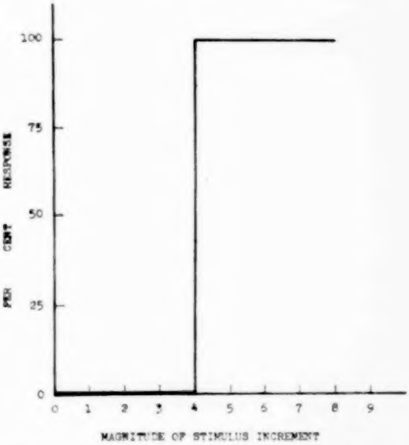

FIG. 1. FORM OF THE PSYCHOMETRIC FUNCTION ON THE SINGLE ASSUMPTION OF THE EXISTENCE OF NEURAL QUANTA

This assumption holds that the

[5] When this assumption is met, the observer is said to have adopted a "one-quantum" criterion of discrimination. Usually, however, for reasons to be indicated in the development of the theory, such a criterion is difficult to establish and the observer will require that the stimulus-increment excite two additional quanta before a discrimination is reported.

over-all sensitivity of the human organgism does not remain at a constant level, but fluctuates momentarily and randomly[6] through magnitudes condierably larger than a single quantum. It follows, therefore, that the amount of stimulus energy required to activate a fixed number of neural units will vary with fluctuations in sensitivity. Conversely, a stimulus of given magnitude will excite a varying number of neural units. Since the variation in the number of activated units is assumed to be quantal, or stepwise discontinuous, all of the available energy of a given stimulus will not necessarily be utilized in a given presentation. Thus, at a particular moment, the given stimulus may excite completely a certain number of quanta and leave a small amount of residual energy which "partially" excites an additional quantum.[7] This residual, while ineffective by itself to activate the next quantum, becomes available for summation with the energy provided in the succeeding stimulus-increment and may consequently produce a discriminatory response.

The basic notions of the theory of the neural quantum which have been introduced up to this point are schematically represented in Fig. 2. Two continua are shown: (*a*) a stimulus continuum with an arbitrary scale, and (*b*) a sensory continuum with discrete neural units. On the stimulus continuum, S is the magnitude of the standard stimulus; ΔSq is the magnitude of the stimulus-increment which will always excite an additional quantum; and ΔS is the amount of energy (magnitude of the stimulus-increment) which is required to activate a "partially" excited neural unit. On the sensory continuum, p is the amount of "partial" excitation resulting from the presentation of a given S.

If Fig. 2 is taken to represent the condition of loudness discrimination, a stimulus magnitude of 17 energy units is considered sufficient to stimulate completely the neural elements: *a*, *b*, and *c*. Neural element "*d*" is only "partially" stimulated by the residual energy beyond 15 units. Assume that such a situation would yield a given loudness. If the stimulus energy were reduced to 15 units, there would be no apparent change in loudness; but, if the energy were reduced to 14 units, neural element "*c*" would drop out and the loudness would diminish by one j.n.d. Likewise, if the energy were increased to 19 units by introducing a 2-unit stimulus-increment, no change in loudness would result. At 20 units, however, the loudness would increase by one j.n.d.

Two features of the diagram in Fig. 2 should be noted specifically: (*a*) the size of the neural quantum is measured in terms of ΔSq, and (*b*) a "partially" excited unit can be stimulated by adding to S an increment (ΔS) smaller than the amount (ΔSq) required for stimulation when no "partial" excitation (p) exists. Fluctuations in sensitivity can, therefore,

[6] This assumption appears in agreement with the available data of Montgomery (**33**) and Lifschitz (**31**) which indicate that the sensitivity of the ear approximates a normal distribution as it varies with time. These fluctuations are presumably due to extraneous factors, such as breathing movements, extraloud heart beats, lapses of attention, shifts in motivation, etc.

[7] This notion of "partial" excitation follows the work of Stevens, Morgan, and Volkmann (**38**). While such a notion is inconsistent with a quantal function, it does aid in the conceptualization of the theoretical model and, hence, will be retained in the present paper. A restatement of the concept in more precise stimulus terms would in no substantial way alter the theory.

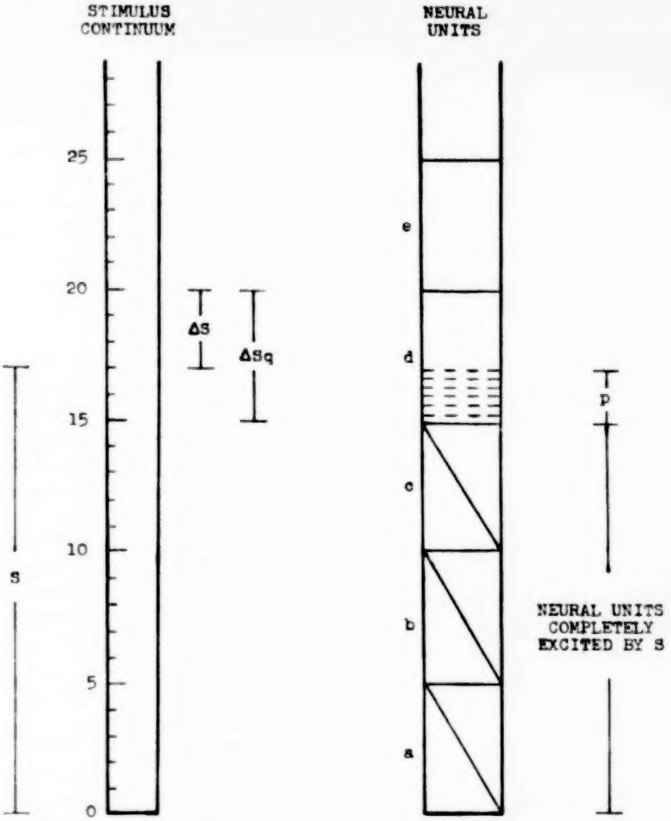

FIG. 2. A SCHEMATIC REPRESENTATION OF THE BASIC NOTIONS INVOLVED IN THE THEORY OF THE NEURAL QUANTUM

bring about discriminatory responses to an increment smaller than one of neural unit size. This would account for the absence of the perpendicular psychometric function (see Fig. 1) predicted on the single assumption of the existence of neural units.

As already indicated, a stimulus-increment (ΔS) smaller than one of neural unit size (ΔSq) will excite an additional quantum only when the residual energy (p) is sufficiently augmented by the increment to provide a total supply of energy equal to or greater than that required for the activation of a "complete" neural unit. Obviously, when the residual is large, the increment required is small; when the residual is small, the increment required is large. Thus, at any instant, the magnitude of the stimulus-increment necessary to add another quantum to the total number excited by the standard stimulus depends upon the amount of residual energy or "partial" excitation. Stated in mathematical form:

$$\Delta S = \Delta Sq - p, \qquad [1]$$

where ΔS is the stimulus-increment required to activate an additional quantum; ΔSq is the size of the increment which will always excite one quantum; and p is the amount of "partial" excitation elicited by the surplus energy in the standard stimulus (S).

Equation 1 indicates that a given ΔS will completely stimulate the additional quantum needed for a discrimination whenever $\Delta S \geqq \Delta Sq - p$. As the size of the increment becomes greater, an increase in the number of discriminations is to be expected. The precise manner in which the number of discriminations increases as a function of increment size depends upon the relative frequency with which the different surplus (residual) values occur.

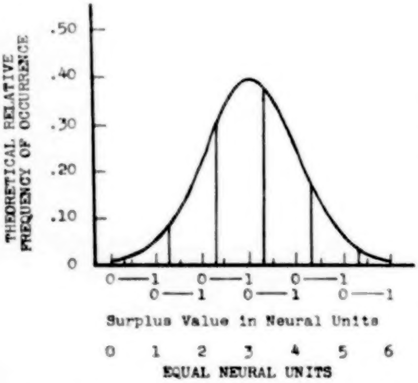

FIG. 3. "CHANCE" DISTRIBUTION OF SURPLUS VALUES ARISING FROM RANDOM FLUCTUATIONS IN ORGANISMIC SENSITIVITY

The relative frequencies of occurrence can be arrived at by the following logical analysis. Assume, as before, that the over-all fluctuation in the sensitivity of the organism is large as compared to the size of a neural quantum. This fluctuation will produce a variation in the number of neural units excited completely by the standard stimulus. Since the amount of surplus or residual energy cannot exceed neural unit size, it must always stimulate "partially" the first unit beyond the last one stimulated by the standard stimulus. The surpluses will be spread out, therefore, over the same range of neural units stimulated during the course of fluctuations in sensitivity. The relative frequency with which the surpluses are distributed over this range will be dependent upon the time distribution of the organism's sensitivity. Assuming, as before, that the organism fluctuates in sensitivity because of a large number of unknown, independent factors, the distribution of surpluses over the range described will approximate a normal curve.

This "chance" distribution of surpluses is shown in Fig. 3. The abscissa has been arbitrarily divided into six equal neural units which represent the range over which the organism fluctuates in sensitivity. Each neural unit has been subdivided, also arbitrarily, into ten equal surplus values. Thus, a surplus of a given magnitude may be found to occur in each of the six neural units. The ordinate of the distribution represents the theoretical relative frequency of occurrence of the surplus values. For example, let the surplus value equal 0.3 of a neural unit. The vertical lines drawn in the distribution will indicate the relative frequency of occurrence of this surplus value within each of the six neural units. Notice that this relative frequency is not the same from unit to unit.

The probability function of the surplus values can now be determined. This is accomplished by summating over the several neural units covered by the normal distribution of surpluses resulting from the organism's fluctuations in sensitivity, the relative frequencies of occurrence for each possible surplus value from zero to neural unit size. From the obtained distribution, the probability of a given surplus value may be determined.

A graphical derivation of the probability function of surplus values can be demonstrated by utilizing the representation in Fig. 3. Accordingly, Fig. 4 shows the function obtained by summating, over the six neural units, the relative frequencies of occurrence of the individual surplus values ranging from zero to one in 0.1 neural unit steps. For example, the segmented vertical line in the diagram of Fig. 4 represents the summation for the surplus value equal to 0.3 of a neural unit. Each segment of the line, starting at the bottom, corresponds in length to the appropriate ordinate shown in Fig. 3. The same procedure has been followed to obtain the summation value for each of the nine remaining surpluses. Since the form of the obtained distribution is approximately rectangular, and the neural unit is divided into ten equal parts, the probability for each surplus is the same. Thus, it may be said that, given a standard stimulus, any surplus value is as likely to occur as any other. A similar conclusion may be arrived at mathematically by Bayes' (**16**) theorem, which states that the distribution of the probability integrals of any continuous curve is a rectangle with every probability between zero and one equally likely.

On the basis of the preceding analysis, the form of the psychometric function may be predicted. It has already been shown that the number of responses to an increment is a function of the size of the increment; the greater the increment, the greater will be the number of responses. The rate at which the number of responses increases with the increase in the size of the increment can be determined from the frequency of occurrence of the surplus values. Since the probability function is rectangular, one value of surplus occurs as frequently as any other. Therefore, for a given increase in the size of the increment, the proportion of surpluses which can be augmented to neural unit size, or greater, is always the same. The rate of increase in the number of responses is, then, constant and the relationship between increment size and percentage of response is clearly linear.

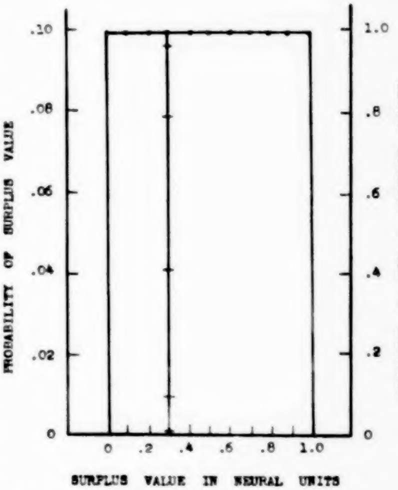

FIG. 4. PROBABILITY FUNCTION OF SURPLUS VALUES

Such a psychometric function may be graphically represented by a straight line, i.e., the integral of the rectangular probability distribution of surplus values. The zero point for this function should correspond to the value of the standard stimulus, since any increment, no matter how small, will find a surplus which it can augment to neural unit size, provided the increment is presented a sufficient number of times. The 100 per cent point should correspond to the smallest increment which always succeeds in exciting an additional neural unit. This increment, which must be independent of the surplus since it always produces a discriminatory response, provides a measure (ΔSq) of the size of the neural quantum.

The foregoing statements of the theory of the neural quantum can be summarized in the form of mathematical equations. Equation 1 has already been formulated ($\Delta S = \Delta Sq - p$) and indicates that an additional quantum will be activated whenever the amount of energy in an increment is sufficient to augment the surplus energy to a neural unit amount. Since the surplus (p) fluctuates between $0 \leqq p \leqq \Delta Sq$ and any value of p is as likely to occur as any other, the proportion of times that an increment will activate an additional neural unit is given by:

$$f_1 = \frac{\Delta S}{\Delta Sq}, \qquad [2]$$

where f_1 is the relative frequency of the instants during which ΔS excites one additional quantum; ΔS and ΔSq have the same meanings as previously given.

Two features of the relationship expressed in Equation 2 should be noted: (*a*) the proportion of responses increases as a linear function of incremental size and (*b*) the value of f_1 may vary from zero to one.

Equations 1 and 2, however, hold only for those discrimination situations in which the excitation of a *single* additional quantum is sufficient to produce a response; but, the evidence of Békésy (**1**), Miller and Garner (**32**), and Blackwell (**4**) shows that usually *two* additional units must be activated before a discriminatory response is reported. This is attributed to the fluctuations in sensitivity which occur during the presentation of the standard stimulus. Since these fluctuations may produce surplus values of neural unit size, the subject finds it difficult to distinguish this excitation from that resulting from the presentation of an adequate increment. If, as indicative of an increment, the subject adopts a certainty criterion which can be met only when two additional quanta are excited, he is then able to distinguish the effect of the surplus alone from the combined effect of increment and surplus. In this case, the proportion of times that a given increment will produce a discriminatory response may be expressed as follows:

$$f_2 = \frac{\Delta S}{\Delta Sq} - 1, \qquad [3]$$

where f_2 is the relative frequency of occurrence of the instants during which ΔS excites two additional quanta. Observe that f_2 may also vary between zero and one.

Equation 3 may be rewritten in terms of the percentage (P) of the increments to which an observer should be able to make a discriminatory response. In this form,

$$P = \left(\frac{\Delta S}{\Delta Sq} - 1\right) \times 100, \qquad [4]$$

and P may vary between 0 per cent and 100 per cent.

Referring to Equation 4, increments less than quantal in size will never stimulate two additional quanta since the surplus cannot exceed one unit; hence, the combined energy of increment and surplus will be less than that required for two units and no discriminatory response will occur. With increments of quantal size or greater, discriminatory responses will occur and will increase in the same manner as described in the case of the "one-quantum" criterion. One hundred per cent response will occur at the smallest increment value which can always excite two additional quantal units. This value, which is independent of surplus, will be twice the size of the largest increment to which a response never occurs. This prediction follows from the assumption previously made that neural units are equal. The largest increment to which a response "just never" occurs is taken as a measure of the size of the first quantal unit; the smallest increment to which a response always occurs is taken as the size of two quantal units. Consequently, on the assumption of equal units, a two-to-one ratio obtains between the value at which the psychometric function reaches 100 per cent and the value at which it first departs from 0 per cent.

If the experimental conditions and underlying assumptions of the quantum theory are satisfied, a typical psychometric function such as that for pitch or loudness discrimination should resemble the function presented in Fig. 5. Two features of the function should be observed: (*a*) there is a linear relationship between the percentage of increments heard and the magnitude of stimulus-increments presented, and (*b*) there is a two-to-one ratio between the values of the function at the 100 per cent point (2 quanta) and the 0 per cent point (one quantum). These features of the psychometric function are the two specific deductions of the neural quantum theory of sensory discrimination which can be subjected to experimental verification.

Quantal Predictions and Techniques of Evaluation

The first major prediction of the theory of the neural quantum is that the percentages of stimulus increments discriminated will be distributed rectilinearly between 0 per cent and 100 per cent. Stated symbolically, P will be a linear function of ΔS. The classical hypothesis, as previously indicated, would predict a sigmoidal probability function for the same set of data. The question becomes, then, "which of the two hypotheses, sigmoidal or quantal, better fits these data points?" (**35**, p. **61**).

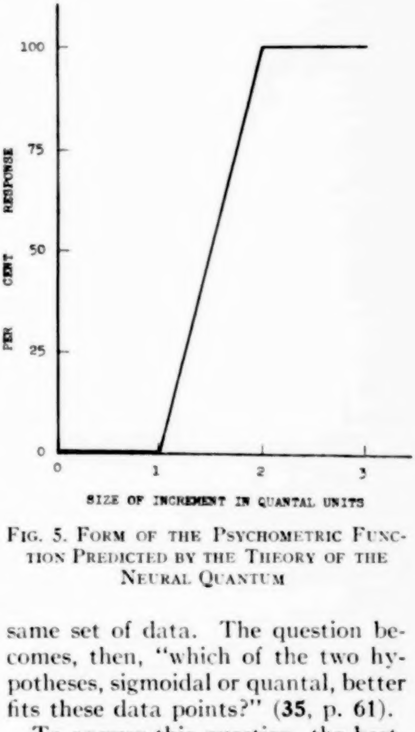

Fig. 5. Form of the Psychometric Function Predicted by the Theory of the Neural Quantum

To answer this question, the best-fitting sigmoidal and rectilinear functions must be constructed for the given set of data.[8] While any one of several techniques may be employed, the curve-fitting process is most ade-

[8] At least one investigator (**11**) has assumed that threshold data may be fitted by a log-Gaussian distribution, i.e., an ogive expressed in terms of a logarithmic scale of stimulus magnitude. Since the normal ogive and log-Gaussian distribution are highly similar, no special case will be made in the present paper for this additional hypothesis.

quately accomplished by using either the method of least squares or the more recently developed technique of probit analysis (**21**). When the best-fitting sigmoidal and rectilinear functions have been obtained, it will usually be found that neither curve gives a perfect fit, i.e., passes through all the data points. Thus, an appropriate statistical test of goodness of fit, such as chi square (**28**), must be applied to determine the probability that the experimental data could have been obtained by chance when the "true" function was either rectilinear or sigmoidal in nature. The results of this analysis will indicate whether the specific theoretical hypothesis being tested should be rejected or retained.

The second major prediction of the quantum theory is that the smallest stimulus-increment at which 100 per cent discrimination occurs will be twice as large as that at which 0 per cent discrimination occurs. This holds only when an observer has adopted a "two-quanta" criterion of judgment.[9] However, regardless of the judgmental criterion adopted, the quantal index may be defined in the general case as follows:

$$QI = \frac{\Delta S_1}{\Delta S_1 - \Delta S_0}, \qquad [5]$$

where QI is the quantal index, or predicted ratio; ΔS_1 is the size of the smallest stimulus-increment at which 100 per cent discrimination occurs; and ΔS_0 is the largest stimulus-increment at which 0 per cent discrimination occurs. Under a "two-quanta" criterion, ΔS_1 always excites two additional quanta, and ΔS_0 always excites one additional quantum.[10] Since the quantal units are assumed to be equal, QI will equal two. For any quantal criterion adopted, Equation [5] will yield an integral value of QI.

In the computation of QI, the values of the stimulus-increments to be used in Equation [5] are obtained by solving, algebraically or graphically, for the 100 per cent and 0 per cent discrimination points in the linear functions fitted to the experimental data.

Some Requirements of the Quantal Method

The demonstration of neural quanta apparently depends upon very rigorous experimental controls. If the relatively large, momentary fluctuations in over-all organismic sensitivity are not to obscure the "true" nature of the discriminatory process, certain precautions must be taken. As stated by Stevens, Morgan, and Volkmann (**38**, p. 319), "we must add ΔI instantaneously, and remove it before the organism is able to change in sensitivity by more than a negligible amount." This requirement dictates certain experimental procedures: (*a*) there must be no time interval between the presentation of the standard stimulus and variable stimulus, and (*b*) the variable stimulus must be of very short duration. If these conditions are not satisfied, the random fluctuations in

[9] While this is the usual criterion adopted by an observer, it is possible for the observer to require that three additional quanta be excited by the stimulus-increment for discrimination to occur. This corresponds to a "three-quanta" criterion of judgment. No discriminations will occur until the increment is sufficient to excite two additional quanta; discriminations will occur 100 per cent of the time when the increment is sufficient to excite three additional quanta. Since this condition does not alter the basic formulation of the quantum theory, it will not be treated independently in the present paper.

[10] In this restricted case, ΔS_0 is the equivalent of ΔSq as previously defined, i.e., it denotes the size of the neural quantum in terms of the stimulus increment which, under a "one-quantum" criterion, would yield 100 per cent discrimination.

over-all sensitivity may be expected to result in nonrectilinear psychometric functions (**38**).

Other requirements have also been specified by Stevens, Morgan, and Volkmann (**38**). (*a*) The observer must experience little difficulty in making discriminatory judgments. This presupposes that the observer is well trained and that the experimental situation maximally aids the focusing of attention and the stabilization of judgment criteria. (*b*) The judgments must be made rapidly enough to eliminate the need for averaging results from different experimental sessions. If possible, all judgments should be made in a single session, thus minimizing the effects of temporal variations. (*c*) Some observers may be aided in directing their attention by introducing a "warning" signal, such as a dim light,[11] at the proper moments in the test trials. This technique enables the observer to adjust to the series of stimulus presentations and serves to reduce the fatigue of sustained attention. (*d*) No transient sounds must be introduced in the transitions between the standard stimulus and the comparison stimulus. If such sounds are present, they may be used as extraneous cues and will tend to distort the resulting psychometric function.

Experimental Tests and Some Critical Comments[12]

Following the earlier work of Békésy (**1**, **2**), Stevens and Volkmann (**37**) tested the hypothesis that loudness discrimination data could be adequately represented by a linear function in accordance with the requirements of the quantum theory. The experimental techniques employed the precautions already outlined in the preceding section of this paper. A single trained observer was used at a frequency of 100 cps presented at five (20, 30, 50, 60, and 80 db) sensation levels (SL, db above threshold). At each SL, the observer listened to a continuous tone whose intensity was increased for 0.15 sec. at 3 sec. intervals. The task of the observer was simply to press a key whenever an increment was heard. Each increment was presented between 50 and 100 times in random blocks of 25 presentations each. The obtained percentages of perceived judgments ranged from 0 per cent to 100 per cent.

Since the obtained psychometric functions showed the predicted rectilinearity and the two-to-one integral relation, it was concluded that the data supported the quantum theory of discrimination. However, two features of the data analysis should be considered: (*a*) the two-to-one integral relation was obtained on the basis of visually fitted psychometric functions, and (*b*) no tests of goodness of fit of these functions were reported.

Stevens, Morgan, and Volkmann (**38**) later extended the preceding study, using six trained observers in pitch discrimination. The procedure employed was essentially the same as in the case of loudness discrimination. A total of 100 judgments was made by each subject at each of several (eight to ten) frequency increments, all of which were less than 10 cps. The standard stimulus was a 1,000 cps tone at 54 db SL presented in random blocks of 25 trials each. Functions were also obtained for a single observer at five—16, 25, 46, 64,

[11] Some observers, however, find the light distracting and prefer to make judgments without this auxiliary cue (**32**, **38**).

[12] These comments are not to be construed as criticisms of individual authors or or journals, but are intended to point out some aspects of the experimental findings or of data analysis which may aid in appraising the present status of the neural quantum theory.

and 90 db—sensation levels, and for another observer at four—25, 30, 54, and 80 db—sensation levels.

In the treatment of data, linear functions were fitted to the experimental values for pitch discrimination by the method of least squares and phi-gamma functions were fitted to the same values by Boring's (**6**) method.[13] For purposes of curve fitting, Δf was taken as the independent variable and all points falling below 3 per cent and above 97 per cent were omitted in computing the constants of the fitted functions. For each of the fitted functions, a chi-square test of goodness of fit was applied and the corresponding P value was determined. For both types of functions, the number of degrees of freedom was taken to be two less than the number of points to be fitted. The results of this analysis showed that in 14 of the 15 sets of data, the P values were higher for the rectilinear functions than for the phi-functions of gamma. In general, the P values for the functions predicted by the quantum theory were above 0.5, whereas those for the "classic" theory were less than 0.5. Furthermore, the two-to-one integral relation was found to hold rather well in most of the 15 sets of data, but the values ranged from 1.89 to 2.34.

While these data obviously favor the quantum theory, a re-examination of Table I in Stevens, Morgan, and Volkmann (**38**, p. 329) shows that the phi-function of gamma, despite yielding generally lower P values, is considered unacceptable in only one of the 15 fits according to Culler's (**12**) interpretation. Nine of the phi-gamma functions have a fit described as "good" or better, compared to 14 rectilinear functions in the same classification. It would appear, therefore, that on the basis of the individual chi-square values obtained in this study both hypotheses remain tenable.

In an attempt to demonstrate a more decisive difference between the goodness of fit for the classical and quantal hypotheses, a composite of the individual P values was also computed by Stevens, Morgan, and Volkmann (**38**). Since the composite P value for all 15 sets of data taken together was 0.931 when the rectilinear functions were fitted and only 0.008 when the phi-gamma functions were fitted, the quantal hypothesis was considered supported and the classical hypothesis was considered quite unacceptable.

Flynn (**17**), however, has made three criticisms of the treatment of data in the Stevens, Morgan, and Volkmann (**38**) study: (*a*) disregarding those points below 3 per cent and 97 per cent was considered unjustifiable when the fitting was done to compare rectilinear and phi-gamma hypotheses since the critical aspects of this comparison involve these extreme values, (*b*) although the observations were weighted for reliability by Urban's[14] method in determining the best-fitting normal ogives, no weighting was reported in fitting the rectilinear functions, and (*c*) $n-3$

[13] This method utilizes the weighting of observations according to Urban's tables (**22, 44**) which contain the products of the Müller weights and Urban weights as they are generally applied in the constant methods. The Müller weights are intended to equalize the effect of the various proportions of judgments upon the determination of the various corresponding values of gamma in fitting observed data to the phi-gamma function; the Urban weights are intended to place greater emphasis on the more reliable observations in solving for the constants of the phi-gamma function by the method of least squares.

[14] See footnote 13.

degrees of freedom (*df*) should have been used for the ogive fit. Flynn (**17**) concludes, nevertheless, that the general finding is probably correct that 14 out of the 15 sets of data are fitted better by a rectilinear function than by a phi-gamma function.

Lewis and Burke (**27**) have also pointed out certain weaknesses in the application of the chi-square test to the Stevens, Morgan, and Volkmann (**38**) data. (*a*) In comparing the goodness of fit of the two different functions, the same quantity was not minimized in the process of obtaining the constants for the fitted functions. In fitting the linear functions, the sum of squared differences between observed and theoretical proportions was minimized; but, in fitting the phi-gamma functions, the sum of the squared differences between observed and theoretical values of gamma was minimized. This would tend to yield chi-square values for the phi-gamma functions that were inexact and probably inflated by an unknown amount. (*b*) In the analysis of the pitch discrimination data for the six observers at 1,000 cps, 54 db SL, four extreme empirical proportions were excluded in determining the constants of the fitted phi-gamma functions, but were included in the calculations of chi square for individual observers. This procedure would also tend to inflate the composite value of chi square. (*c*) There were seven theoretical proportions representing small theoretical frequencies (less than 10) which should preferably have been combined with adjacent proportions. When these factors were considered in the calculation of new values of chi square for the phi-gamma functions of the six observers, the composite chi square was 11.76, with 10 *df*, as compared to the original value of 33.80, with 21 *df*.[15] Since the recalculated value of chi square falls at about the 30 per cent level of confidence, the phi-gamma hypothesis cannot be rejected.

In a study on pitch discrimination, Flynn (**17**) used three trained observers to listen to a continuous 1,000 cps tone at a 55 db SL. The tone, lasting 1.25 min. per block of 25 trials, periodically changed in frequency for 0.30 sec. The task of the observer was to report each time this change in frequency was detected. The number of trials for each increment was usually 50 or 100, with extremes of 35 and 200. Thirty sets of data were obtained from the three observers. For each set of data, the best-fitting normal ogive and the best-fitting straight line were computed by the method of least squares. Precautions were taken to weight the empirical proportions in fitting both of these functions so that any observed differences in goodness of fit could not be attributed to differences in fitting techniques.[16]

On the basis of chi-square tests of goodness of fit as adapted by Thomson (**40**),[17] 16 sets of data were found to fit neither a straight line nor a

[15] The computations for the chi-square values of the linear functions had the same weaknesses as those outlined for the phi-gamma functions, except that none of the empirical proportions omitted during the curve-fitting process were later included in the tests of goodness of fit. Lewis and Burke (**27**) also mention the fact that the linear functions were not obtained from weighted proportions.

[16] For the normal ogive the Müller-Urban weights, corrected for the number of observations, were used; for the straight line, only the Urban weights were needed and applied.

[17] There were two exceptions to Thomson's (**40**) procedure: (*a*) no attempt was made to compare chi-square values by means of the standard error, and (*b*) the number of *df* for the ogive was the number of percentages minus 3; for the straight line, minus 2.

normal ogive. Four sets of data were found to fit a normal ogive better than a straight line, while the remaining ten sets of data could be fitted better with a straight line, than with a normal ogive. However, some of the weaknesses in data analysis mentioned by Lewis and Burke (**27**) were also evident in this study, e.g., the 0 per cent and 100 per cent points were omitted in determining the best fitting curves but were reintroduced in testing for goodness of fit. In addition, the two-to-one criterion demanded by the quantum theory did not hold in those cases in which linear functions were obtained. Thus, it would appear that the evidence of this study contrary to Flynn's (**17**) interpretation, may be considered as failing to support the neural quantum theory since both predictions were not met.

Koester and Schoenfeld (**26**), while comparing the relative merits of quantal and nonquantal procedures in pitch discrimination, were also interested in duplicating the quantal findings. In the quantal procedure, two highly trained observers were presented with a moderately loud 1,000 cps tone. The standard was given for 1.0 sec. followed without interruption by an increment of the same duration. Twenty standard-increment pairs and two pairs in which the standard continued for 2.0 sec. comprised a series of trials. The same increment was used throughout a series. The task of the observer was to report each time an increment was heard.

For each observer, a complete set of psychometric data was obtained by the quantal method on each of four days. It was concluded that none of the data exhibited either the rectilinearity or the integral relation predicted by the quantum theory. However, no mention was made of fitted functions or tests of goodness of fit. Also, since each of the eight psychometric functions was based on either six or seven points with only twenty observations per point, the conclusions of this study should be accepted with caution.

In a study designed to investigate some of the factors which might obscure the quantal nature of the discriminatory mechanism, Miller and Garner (**32**) obtained intensity discrimination data for a 1,000 cps tone at 40 db SL on two observers using both the standard quantal procedure of Stevens, Morgan, and Volkmann (**38**) and a modified quantal procedure. In the modified procedure, the stimulus-increment was altered at random after each presentation and the observer was not permitted to stop after every 25 presentations. This modification was introduced to prevent the observer from establishing a fixed "two-quanta" criterion of judgment.

The results obtained by the standard quantal method showed that the two psychometric functions could be adequately represented by a straight line fitted by the method of least squares and that the predicted integral relation was closely approximated. For the modified procedure, the phi-gamma functions were fitted by the technique proposed by Guilford (**22**); the quantal hypothesis was evaluated by fitting a series of three straight lines to the empirical values lying between the successive quantal points as determined by the previously-administered standard method. On the basis of chi-square tests of goodness of fit, it was concluded that the quantal hypothesis provided a better description of the data than did the phi-gamma hypothesis.

While these findings are indeed significant, two aspects of data analysis should be pointed out: (*a*) the question arises as to whether data assumed to have been obtained under three different certainty criteria and fitted by three different linear functions can be validly "connected" to yield a *single* psychometric function; and (*b*) assuming that a single function is valid, each of the straight lines at the two extremes of the function must be made to intersect an additional horizontally-linear segment if the function is not to predict impossible response values greater than 100 per cent and less than 0 per cent.

In further analyzing the data of Stevens and Volkmann (**37**) and of Flynn (**17**), Miller and Garner (**32**) proceeded to show that (*a*) the proposed technique of fitting three linear segments to psychometric functions holds in general for those cases involving criterion shifts by the observer and is not limited to loudness discrimination or to the random method of stimulus presentation, and that (*b*) combining of data obtained either under different experimental conditions or from different observers tended to yield psychometric functions in accordance with the phi-gamma hypothesis. Thus, the work of Miller and Garner (**32**) tends to support the quantum theory and serves to isolate some of the factors responsible for nonquantal findings.

In a fairly extensive study, Corso (**10**) recently attempted to test the hypothesis that the data obtained in the auditory discrimination of frequency and intensity satisfied the conditions predicted by the theory of the neural quantum. For intensity discrimination, the general procedure followed that of Stevens and Volkmann (**37**); for frequency discrimination, that of Stevens, Morgan, and Volkmann (**38**). In all, 20 subjects were used in the study, after having been screened from a larger group of 45 by means of an audiometric test and the Seashore pitch and loudness tests. Each subject was given two separate practice hours (for a total of 425 to 1,225 judgments) under the specific conditions of the test trials. Five subjects were tested under each of the following conditions of frequency discrimination: (*a*) 1,000 cps at 20, 40, 60, and 80 db SL, and (*b*) 300, 1,000, and 3,000 cps at 60 db SL. A similar pattern was followed for intensity discrimination. For each test condition, at least six stimulus increments were presented, with approximately 200 judgments being made at each increment-value.

In the analysis of data, linear functions were fitted to the individual sets of frequency and intensity discrimination data by the method of least squares, with all empirical proportions greater than 0.97 or less than 0.03 omitted. The chi-square test was used to test the goodness of fit of each obtained function. In the calculation of chi-square values, all theoretical proportions greater than 0.97 and less than 0.03 were appropriately combined with adjacent proportions. This technique insured that in calculating the chi-square values (*a*) no proportions representing theoretical frequencies of less than five were used, and (*b*) no empirical proportions were used which did not enter into the solutions of the parameters of the linear functions. Of the 70 chi-square values computed, only nine (seven in frequency discrimination, two in intensity discrimination) had a P value equal to or greater than 0.05. Of the nine psychometric functions in which the hypothesis of linearity was retained, only one had a ratio-value which approached the predicted two-to-one criterion. It was concluded, therefore, that the experimental results in frequency and in intensity discrimination failed to satisfy the predictions of the theory of the neural quantum.

Licklider (**29**), in reviewing Corso's (**10**) study, pointed out that failing to obtain psychometric functions which conform to quantal predictions can only disprove the quantum theory if "all the non-physiological error variance" has been eliminated from the experimental measurements and the observer is worked at his "physiological limit." Obviously, the existence or nonexistence of these qualifying conditions can (perhaps) never be known, but only assumed from the obtained data. One would expect, however, that in two essentially identical experiments such a source of error would be roughly equivalent, unless some unusual (and perhaps drastic) precautions were taken in one experiment and not the other.

The theory of the neural quantum has been extended to include the problem of sensory discrimination in areas other than audition. Jerome (**25**) obtained olfactory psychometric functions using stimulus pressure, as measured by an Elsberg olfactometer, as the independent variable. In the discrimination tests, one trained and one untrained observer, after becoming acquainted with the odor of citral, were instructed to indicate by their responses whether or not the odor was present when the stimulus was delivered. The task was presented as one of distinguishing between the stimuli from a control bottle and those from a citral bottle. There were ten presentations of the stimulus from each bottle at each of several (seven to nine) pressure values.

Seven psychometric functions were obtained from the data of the two observers in the discrimination experiment, and six psychometric functions were obtained on two observers in the preliminary test of instructions. Linear functions were fitted to each of these 13 sets of data by the method of averages. The results of this analysis showed that the criterion of rectilinearity was satisfactorily met;[18] but, since the additional criterion of the integral relation was not met, it was concluded that the existence of a differential olfactory quantum was not demonstrated.

There are three apparent weaknesses in Jerome's (**25**) study if the data are to be used to evaluate the quantum theory: (*a*) a nonstandard quantal procedure was used inasmuch as an interval of 30 sec. was permitted to elapse between stimulus presentations to avoid olfactory fatigue; (*b*) only ten observations were made at the critical values at the extremes of the psychometric functions,[19] and (*c*) no tests of goodness of fit were reported, presumably due to the presence of small theoretical frequencies which precluded the use of chi square. Thus, it appears that the data of this study cannot form an adequate basis either for the acceptance or for the rejection of the quantum theory.

DeCillis (**14**) attempted to follow the quantal procedure in finding the relation between amplitude of stimulus movement over a cutaneous area and frequency of positive response.

[18] Since no tests of goodness of fit were reported, it is presumed that rectilinearity was determined by visual inspection of the fitted functions.

[19] For example, of the 13 functions obtained five had no observations at stimulus values yielding between 80 per cent and 100 per cent response; seven had no observations between 0 per cent and 20 per cent response.

The procedure employed was to present a fine air column, at a pressure of 35 lbs. per sq. in., at a point on the skin for 0.10 sec. The air column then traveled across the skin at a rate of 143 mm./sec. to another point where it was again stationary for at least 0.10 sec. After the air column was turned off, the needle controlling the stimulus returned to its starting position. This procedure was repeated 20 times in a series, with the same amplitude of movement presented on each trial. Three subjects were used and sensitivity was measured on the fingertip, arm, and leg. The task of the observer was to report "yes" whenever movement of the air column was perceived and "no" whenever it was not.

The best-fitting straight lines were calculated by the method of least squares for 35 selected psychometric functions, with the 0 per cent and 100 per cent points omitted in the curve-fitting process. The chi-square test for goodness of fit was applied following the method of Brown and Thompson (**9**). No attempt was made to fit 16 sets of nonhomogeneous data. The results of this analysis yielded 20 chi-square values with probability values equal to or greater than 0.95, while 15 values were smaller than this. It was concluded, therefore, that it was not "unreasonable to maintain that the best-fitting psychometric function is rectilinear" (**14**, p. 47). However, in those cases where the hypothesis of linearity was retained, the criterion of the integral relation did not hold. DeCillis (**14**, p. 49) contends that "apparently the integral relation is not to be expected in studies of absolute sensitivity."

It is unfortunate that in this study (**14**) extensive data were not collected at those points on the psychometric functions where maximal differences between the phi-gamma and quantal hypotheses were to be expected. In the 20 out of 35 fitted functions in which the linearity hypothesis was retained, 19 functions had no observations at stimulus values yielding between 0 per cent and 15 per cent (or more) responses; 10 functions had no observations between 85 per cent (or less) and 100 per cent responses. Thus, for a given set of data, since the 0 per cent and 100 per cent points were also omitted in the curve-fitting process, the remaining empirical points would probably not have deviated from a straight line, whether or not the "true" function were ogival or linear.

In the most recent attempt to evaluate the quantal hypothesis, Blackwell (**4**) obtained visual discrimination data for four observers using normal binocular viewing and natural pupils at a luminance of 4.71 foot-lamberts. The stimulus was a circular luminance-increment, subtending 18.5′ located 7° to the right of the fixation spot and was presented for a duration of 0.06 sec. once every 12.25 sec. Each psychometric function was based on 14 to 18 increment-values with 20 observations at each point. The increments were presented in random blocks of 20 trials each and the same increment was used throughout all the trials of a block. The observer indicated discrimination by responding "yes" or "no" to each presentation of the stimulus.

In the data analysis, a linear function was fitted to a selected set of data for each observer.[20] The selection of a specific set of data from

[20] The exact method of curve fitting is not specified in Blackwell (**4**), but the use of probit analysis is implied.

among the total number of sets available (Observer 1: 4 sets; 2, 18 sets; 3, 24 sets; and 4, 10 sets) was carried out in such a way as to obtain the function most adequately fitted by a quantal curve. The results of this analysis revealed that the data of two observers could not be used to evaluate the quantal hypothesis. For one of these subjects, the stimuli were spaced too widely over the critical psychometric range; for the other, the data were extremely scattered. The data of the remaining two subjects were fitted adequately in most cases by either a "two-quanta" or a "three-quanta" curve.

Blackwell (**4**), however, considers this apparent confirmation of the theory as spurious. This assertion is based upon the fact that, as the criterion of judgment increases from two quanta to three quanta, the 50 per cent threshold decreases rather than increases as expected from an extension of the quantum theory. The hypothesis is advanced that "response channelization" (**5**) may actually be responsible for the fact that some experimental data appear to conform to quantal rectilinearity. It is concluded (**4**) that the visual threshold-data obtained by the standard quantal procedure do not confirm the predictions of the neural quantum theory.

Discussion

While the predictions of the theory of the neural quantum are specific: (*a*) rectilinearity of the psychometric function, and (*b*) an integral relation between the values of the stimulus-increments at the 100 per cent and 0 per cent response points on the psychometric function, the experimental task to evaluate these predictions is extremely difficult. As Blackwell (**4**, p. 398) states, "Essentially, the quantum theorists have so restricted the allowable conditions of measurement and the analysis of data that it is difficult to obtain an unambiguous evaluation of the theory." Licklider (**29**, p. 99) has aptly summarized the difficulties to be encountered in attempting to obtain negative evidence on the theory of the neural quantum by saying that "it is a shame that the quantum theory has such strong built-in self-protection."

In the first place, the applicability of the theory is restricted to data collected under one psychophysical procedure: the quantal method (**21**). This is unlike the usual approach to psychophysical research where one or more methods may be appropriate for the investigation of a given problem. Within the quantal method, Blackwell (**4**) objects to the "phenomenal report" as the only "indicator-response" and cites data (**3**) to support his contention that the "forced-choice" technique[21] is more adequate than the phenomenal report in threshold measurements under routine conditions. It is maintained that the use of "forced-choice" as the "indicator-response" tends to minimize session-to-session variability when practiced observers are used.

A second restriction placed on the data-collection process is that stimulus increments must be grouped into blocks of presentations of the same magnitude. Miller and Garner (**32**) have demonstrated clearly that the random ordering of stimulus magnitudes prevents even the well-trained

[21] The "forced-choice" technique is defined by two conditions: (*a*) the observer "is required to indicate discrimination by correctly identifying some verifiable attribute of the stimulus such as its spatial location or temporal interval; and (*b*) he is required to select an answer on each stimulus-presentation—even if he has to guess" (**4**, p. 398).

observer from adopting a stable criterion, but the interpretation of the single(?) function obtained is not too clear. Blackwell (**4**) argues that the nonrandom block presentations provide the observer with the opportunity to respond in an invalid manner. It is maintained that the presence of "positive response channelization" and of "negative response channelization" will operate to distort threshold data into a form resembling that required by the quantum theory.[22] Senders and Sowards (**36**), in a study in which judgments were made of the simultaneity of presentation of a light and a tone, also found that successive presentations of the stimulus near threshold tended to produce long series of identical responses.

Koester and Schoenfeld (**26**, p. **11**) have a view somewhat similar to Blackwell's (**4**), contending that in the course of prolonged practice the observers may "learn to adjust and cut-and-fit their certainty criteria to the several values of the comparison stimuli in such a way as to yield the necessary rectilinearity in the psychometric function and the integral relation."

Osgood (**35**, p. 64) raises the same question on methodology by asking: "if the subject knows that all increments in a given series are going to be identical, is it not possible for him to set up a *subjective standard* on the basis of which he graduates the frequency of his responses?" Some evidence on this point is reported in the study of Senders and Sowards (**36**) in which it was found that the observers tended to adjust their proportion of responses in accordance with their expectations.

A third restriction which must operate in quantal studies is that only data collected in a single experimental session may be used to test the quantal predictions. Miller and Garner (**32**) have demonstrated that two sets of data obtained from the same observer by the same procedure but at different times will average into a typical sigmoidal distribution, even though the separate functions are rectilinear. Nevertheless, such a restriction makes it practically impossible to establish the adequacy of the quantum theory with any high degree of confidence. Through a special application of the chi-square test, Blackwell (**4**) has shown that no matter how many sets of experimental data are available, at least 40 presentations must be made in each experimental session at each of 10 stimulus-values if the normal ogive is to fail to fit the data, even though the data may actually conform to the specific requirements of the quantum theory. In the studies reviewed in the present paper, most functions were based on less than 10 stimulus-values each and many functions utilized 25 or less observations per experimental point. Apparently, under these conditions, unequivocal results could not have been expected. An additional comment is made by Lewis and Burke (**27**) who indicate that, in data analysis, the elimination of extreme proportions makes it impossible to obtain a critical evaluation of the phi-gamma hypothesis through the use of the chi-square test of goodness of fit.

There is some evidence to show that, perhaps, the restriction on the

[22] "Positive response channelization" is defined as an increase in the number of perceived judgments toward the ends of the blocks of stimuli for which the predominant response was positive. "Negative response channelization" is defined as an increase in the number of nonperceived judgments toward the ends of the blocks of stimuli for which the predominant response was negative.

combining or averaging of experimental data is unjustified. Koester and Schoenfeld (**26**) have shown that threshold data remain fairly consistent from day to day; Corso (**10**) found no differences among data collected in two successive hours; Myers and Harris (**34**) have reported that the fluctuation of the auditory threshold is approximately less than 1 db for relatively short periods of time. It should be recalled, however, that the assumption of momentary and random changes in the over-all sensitivity of the observer is fundamental to the derivation of the quantum theory. Thus, any compromise on this restriction would undoubtedly necessitate a major revision of the theory. Nevertheless, since the integral relation predicted on the assumption of relatively large fluctuations in sensitivity has seldom been obtained, it may be that in the final analysis this assumption may prove to be unwarranted.

In addition to the methodological problems already mentioned, certain other issues have been raised. Osgood (**35**) is disturbed by the observation reported in Stevens, Morgan, and Volkmann (**38**, p. 334) that some of the stimulus-increments are perceived as "larger and plainer than others. Increments heard 80% of the time tend to be subjectively larger than increments heard only 20% of the time." If, as developed in the quantum theory, "discrimination depends upon the addition of another neural unit to those already in operation, how can the *same* additional quantal unit seem smaller when it is added only 20% of the time as compared with its being added 80% of the time?" (**35**, p. 65). While this might well be a critical issue for the theory of the neural quantum, its resolution is not readily apparent, despite Jerome's (**25**) contention that evidence was obtained for the existence of a "magnitude quantum" of olfactory sensitivity.

Wever (**46**) has taken the position that *peripheral* quanta have not, as yet, been demonstrated and are not to be expected on the basis of the volley theory of hearing. According to the volley theory, pitch is considered to be a continuous function of volley frequency for low and intermediate tones, and of spatial pattern for high tones; loudness is considered to be a continuous function of the magnitude of auditory nerve discharge which depends upon the number of active fibers and the rates of fiber activity.

It should be recalled, however, that Stevens, Morgan, and Volkmann (**38**) hypothesize that (*a*) the neural quantum appears at a *central* not a *peripheral* locus, (*b*) it is functional rather than anatomical, and (*c*) it involves a number of fibers rather than single fiber. Three arguments are offered to support this view as opposed to Békésy's (**1**) contention that the quantal unit is the individual nerve fiber: (*a*) the quantum for the individual observer has no fixed magnitude, (*b*) for a given sensory attribute, the number of auditory nerve fibers is greater than the number of quanta, and (*c*) the binaural quantum is approximately two-thirds the size of the monaural quantum. It should also be realized that the neural units of Stevens, Morgan, and Volkmann (**38**), whether or not substantiated, are hypothetical constructs and do not specify the neural correlates of sensory attributes (**35**).

One final point remains to be considered. The derivation of the theory of the neural quantum is based on the assumption of two quantitative variables: (*a*) a physical continuum and (*b*) a psychological continuum.[23] The modern theory of psychophysics, however, assumes three parallel quantitative variables: (*a*) a physical continuum, (*b*) a sensory response continuum and (*c*) a judgment continuum (**21**). According to this schema, the notions of neural quanta are most directly related to continuum (*b*). However, there is a considerable amount of experimental evidence which shows that the regression relating the judgment continuum (*c*) to the sensory response continuum (*b*) is not always linear, and neither is the correlation always perfect. Thus, in the psychometric function, which relates continuum (*c*) to continuum (*a*), it would be possible to obtain a curve unlike that which relates continuum (*b*) to continuum (*a*). In other words, quantal functioning at continuum (*b*) might not be evidenced in a psychometric function unless a perfect, linear relationship between (*c*) and (*b*) could be obtained under certain conditions of careful experimental controls, attentive attitude of the observer, stabilized learning, etc. On the other hand, it is conceivable that quantal functioning might characterize continuum (*c*) independently of the quantal or nonquantal character of continuum (*b*). Thus, even if the existence of a rectilinear psychometric function were to be unequivocally established, the interpretation of the processes underlying the function would not be immediately apparent.

Summary and Conclusions

The present paper has attempted to fulfill two primary objectives: (*a*) to present a complete and detailed account of the theory of the neural quantum of sensory discrimination, and (*b*) to review the literature on the quantum theory in order to assess the current status of the theory in the light of the total experimental evidence now available.

It has been shown that from the theory of the neural quantum two specific hypotheses about the psychometric function may be derived and tested: (*a*) that a linear relationship obtains between the size of stimulus-increments presented and the percentage of responses observed, and (*b*) that an integral relation obtains between the stimulus-increment values of the function at the 100 per cent and 0 per cent points-of-response. Data in support of these hypotheses would indicate that the fundamental processes involved in sensory discrimination are discrete or quantal in character.

While the hypotheses derived from the quantum theory are experimentally verifiable, severe limitations in methodology and in statistical treatment of data make it extremely difficult to evaluate the tenability of the hypotheses as opposed to the alternate views of the phi-gamma function. However, despite these limitations, it may be concluded that in certain investigations rectilinear psychometric functions have been obtained. The existence of the integral relation, contrariwise, has seldom been demonstrated. Thus, when both factors are considered in the body of available evidence, it appears that unequivocal support of the neural quantum theory is, for the most part, lacking. In addition, the validity of judgments obtained under the experimental conditions of the quantal method has been seriously questioned.

The present review of literature on

[23] The term "continuum" is considered in a broad sense and permits a quantum theory for either variable.

the quantum theory suggests the need for future research along two major lines: (*a*) the development of a more satisfactory technique for statistically testing the goodness of fit of the quantal and phi-gamma hypotheses to a set of experimental data, and (*b*) the determination of the specific conditions under which rectilinear psychometric functions may be obtained in order to establish the validity and universality of quantal notions. Until such research is carried out, the issue of the neural quantum theory of sensory discrimination cannot be fully resolved.

REFERENCES

1. Békésy, G. V. Über das Fechner'sche Gesetz und seine Bedeutung für die Theorie der akustischen Beobachtungsfehler und die Theorie des Hörens. *Ann. Physik*, 1930, **7**, 329–359.
2. Békésy, G. V. Über die Hörschwelle und Fühlgrenze langsamer sinusförmiger Luftdruckschwankungen. *Ann. Physik*, 1936, **26**, 554–566.
3. Blackwell, H. R. Psychophysical thresholds: Experimental studies of methods of measurement. *Engng. Res. Inst. Bull.*, No. 36. Ann Arbor: Univer. of Michigan Press, 1952.
4. Blackwell, H. R. Evaluation of the neural quantum theory in vision. *Amer. J. Psychol.*, 1953, **66**, 397–408.
5. Blackwell, H. R. The influence of data collection procedures upon psychophysical measurement of two sensory functions. *J. exp. Psychol.*, 1952, **44**, 306–315.
6. Boring, E. G. Urban's tables and method of constant stimuli. *Amer. J. Psychol.*, 1917, **28**, 280–293.
7. Boring, E. G. A chart of the psychometric function. *Amer. J. Psychol.*, 1917, **28**, 465–470.
8. Boring. E. G. Auditory theory with special reference to intensity, volume, and localization. *Amer. J. Psychol.*, 1926, **37**, 157–188.
9. Brown, W., & Thomson, G. H. *The essentials of mental measurement.* (4th ed.) Cambridge: Cambridge Univer. Press, 1940.
10. Corso, J. F. The neural quantum in discrimination of pitch and loudness. *Amer. J. Psychol.*, 1951, **64**, 350–368.
11. Crozier, W. J. On the visibility of radiation at the human fovea. *J. gen. Physiol.*, 1950, **34**, 87–136.
12. Culler, E. Studies in psychometric theory. *J. exp. Psychol.*, 1926, **9**, 169–194.
13. Culler, E. Studies in psychometric theory. *Psychol. Monogr.*, 1926, **35**, No. 2 (Whole No. 163). Pp. 56–137.
14. DeCillis, O. E. Absolute thresholds for the perception of tactual movement. *Arch. Psychol.*, 1944, **294**, 1–52.
15. Fechner, G. T. *Elemente der psychophysik.* Leipzig: Breitkopf und Härtel, 1860.
16. Fisher, A. *The mathematical theory of probabilities.* New York: McGraw-Hill, 1936.
17. Flynn, B. Pitch discrimination. *Arch. Psychol.*, 1943, **280**, 3–41.
18. Fullerton, G. S., & Cattell, J. M. On the perception of small differences. *Univer. Pennsylvania Philos. Series*, No. 2, 1892.
19. Gabor, D. Acoustical quanta and the theory of hearing. *Nature*, 1947, **159**, 591–594.
20. Granit, R. *Sensory mechanisms of the retina.* London: Oxford Univer. Press, 1947.
21. Guilford, J. P. *Psychometric methods.* (2nd Ed.) New York: McGraw-Hill, 1954.
22. Guilford, J. P. *Psychometric methods.* (1st Ed.) New York: McGraw-Hill, 1936.
23. Hecht, S., Shlaer, S., & Pirenne, M. H. Energy, quanta, and vision. *J. gen. Physiol.*, 1942, **25**, 819–840.
24. Helmholtz, H. F. L., *On the sensations of tone.* New York: Longmans-Green, 1912.
25. Jerome, E. A. Olfactory thresholds measured in terms of stimulus pressure and volume. *Arch. Psychol.*, 1942, **274**, 1–44.
26. Koester, T., & Schoenfeld, W. N. Some comparative data on differential pitch sensitivity under quantal and non-quantal conditions. *J. gen. Psychol.*, 1947, **36**, 107–112.
27. Lewis, D., & Burke, C. J. The use and misuse of the chi-square test. *Psychol. Bull.*, 1949, **46**, 433–489.
28. Lewis, D. *Quantitative methods in psy-*

chology. Ann Arbor: Edwards Bros., 1948.

29. Licklider, J. C. R. Hearing. *Annu. Rev. Psychol.*, 1953, **4**, 89–110.

30. Licklider, J. C. R. Basic correlates of the auditory stimulus. In S. S. Stevens (Ed.) *Handbook of experimental psychology*. New York: Wiley, 1951.

31. Lifschitz, S. Fluctuations of the hearing threshold. *J. acoust. Soc. Amer.*, 1939, **11**, 118–121.

32. Miller, G. A., & Garner, W. R. Effect of random presentation on the psychometric function: implications for a quantal theory of discrimination. *Amer. J. Psychol.*, 1944, **57**, 451–467.

33. Montgomery, H. C. Influence of experimental technique on the measurement of differential intensity sensitivity of the ear. *J. acoust. Soc. Amer.*, 1935, **7**, 39–43.

34. Myers, C. K., & Harris, D. J. The inherent stability of auditory threshold, *U. S. Navy Bur. Med. Surg. Rep.*, April, 1949, Proj. NM-003-021.

35. Osgood, C. E. *Method and theory in experimental psychology*. New York: Oxford Univer. Press, 1953.

36. Senders, V. L., & Sowards, A. Analysis of response sequences in the setting of a psychophysical experiment. *Amer. J. Psychol.*, 1952, **65**, 358–374.

37. Stevens, S. S., & Volkmann, J. The quantum of sensory discrimination. *Science*, 1940, **92**, 583–585.

38. Stevens, S. S., Morgan, C. T., & Volkmann, J. Theory of the neural quantum in the discrimination of loudness and pitch. *Amer. J. Psychol.*, 1941, **54**, 315–335.

39. Thomson, G. H. Accuracy of the $\phi(\gamma)$ hypothesis. *Brit. J. Psychol.*, 1914, **7**, 44–55.

40. Thomson, G. H. The criterion of goodness of fit of psychophysical curves. *Biometrika*, 1919, **12**, 216–230.

41. Thurstone, L. L. The phi-gamma hypothesis. *J. exp. Psychol.*, 1928, **11**, 293–305.

42. Titchener, E. B. *A textbook of psychology*. New York: Macmillan, 1919.

43. Troland, L. T. *Psychophysiology*. Vol. II. *Sensation*. New York: Van Nostrand, 1930.

44. Urban, F. M. Die psychophysischen Massmethoden als Grundlagen empirischer Messungen. *Arch. ges. Psychol.*, 1909, **15**, 261–415.

45. Volkmann, J. Quantum theory in psychology. *Trans. N. Y. Acad. Sci.*, 1941, **3**, 213–217.

46. Wever, E. G. *Theory of hearing*. New York: Wiley, 1949.

Received November 23, 1955.

PSYCHOLOGICAL BULLETIN
Vol. 53, No. 5, 1956

TECHNIQUES FOR COMPUTING SHIFT IN A SCALE OF ABSOLUTE JUDGMENT[1]

KURT SALZINGER
Columbia University[2]

The method of absolute judgment (single stimuli) has a long history. It started originally as a psychophysical method, i.e., it was applied to physical stimuli. Since then, however, it has come to be more widely applied, i.e., to stimuli which cannot easily be ordered on a physical continuum.

McGarvey describes the method of single stimuli in the following way: "The observer is simply presented with the members of a group of stimuli one at a time and asked to render a judgment upon each by assigning it to one of a specified set of categories" (**7**, p. 9). This assignment of stimuli to categories is sometimes referred to as a "naming response."

Investigators using the method of absolute judgment have referred to the observer's behavior as a formation of a frame of reference (see Helson, **4**, for example) in accordance with which he responds to each of the stimuli which he must judge. Experimenters were also interested in discovering how to modify the frame of reference of an observer. This modification of the frame of reference is known as shift and has been brought about by, among others, such variables as anchor stimuli (stimuli not originally included among the stimuli presented to *S*) and social stimuli (the judgments rendered by another *S*).

It may be defined as a systematic change in the categories to which the observer assigns the members of a group of stimuli or as a systematic change in the stimulus values to which he gives particular naming responses.

Along with the study of the parameters affecting absolute judgment have come a number of different techniques for the statistical treatment of the data to arrive at a measure of shift. In this paper, an attempt will be made to review the techniques used up to the present time for computing the amount of shift as well as to present two new techniques.

Perhaps the most commonly used technique for measuring shift is based upon the use of ratings. In this method numbers are applied to the categories either during the experimental situation or afterwards and these numbers are then treated as an equal ratio scale. To evaluate whether a shift in judgment has taken place from one condition to another, means, differences between means, variances, and critical ratios of these ratings are calculated. The technique of ratings has been used by, among others, Helson (**4**) for judgment of

[1] This article is based on a dissertation submitted in partial fulfillment of the requirements for the Ph.D. degree in the Department of Psychology in the Faculty of Pure Science, Columbia University. The author is indebted to Professors J. Zubin, H. Garrett, and C. G. Mueller, Dr. S. Kugelmass, and Mrs. S. Salzinger for their invaluable aid during the various phases of the dissertation on which this paper is based. He also wishes to acknowledge his gratitude to Professor J. Zubin, Dr. E. Burdock, Dr. S. Sutton and Mrs. S. Salzinger for their comments on this paper. The preparation of this article was facilitated in part by grant M586 from NIMH.

[2] Now at Biometrics Research, N. Y. State Dept. of Mental Hygiene, Psychiatric Institute, 722 West 168 St., New York 32, N. Y.

weights as well as in the field of vision. Both Heintz (**3**) and Brown (**1**) apparently felt the need for justifying the use of this method. The former did so by finding fault with other methods of data treatment (in which incidently he was justified), and the latter justified the procedure by appealing to the fact that Likert (**5**) found high positive correlations between scores based upon frequency counts of judgments and arbitrary ratings applied to the attitude items of a questionnaire. Brown ignores the fact that Likert used judgments of different stimuli, i.e., Brown used weights while Likert used verbal stimuli. Furthermore, the fact remains clear that this method of treating data consists of the application of numbers to events (judgments in this case) without specifying the operations in the experimental situation that would be equivalent to the operations involved when the numbers are combined statistically. It is here suggested that this method cannot be used since the operations performed with the numbers cannot be performed with the events (judgments) being quantified. In other words, no evidence is available to show that a rating of "4" is two times as great as one of "2," etc.

Likert (**5**) uses ratings indirectly. He bases the values he assigns to judgments on the following: he assumes that the judgments are normally distributed; then he determines the value of each category by converting the proportion of *S*s giving each judgment (or the proportion of responses given by one *S*) to a standard score, which in turn is based on the assumption that the use of standard deviations results in equal interval scales of the judgment continuum. This method is somewhat long and necessitates a large number of judgments or a large number of judges to obtain reliable proportions. In cases where one postulates individual differences, it becomes necessary to make separate calculations of the numerical values of each category of judgment for each *S*. Such a procedure makes the already long procedure longer. Furthermore, the assumption of normality cannot always be justified or met.

A third method which makes use of only the assumption of an ordinal scale was utilized by Mausner (**6**). He took median values of each *S*'s judgments in two different situations. He then plotted *Ss*' median judgments for groups of 20 trials, showing graphically what *S*s shifted and under what conditions. While he did not apply any statistical tests to these scores (he did to different types of scores to be discussed below), this type of data lends itself to easy analysis by means of nonparametric tests like the median test, described in Mood (**8**). The point might be made here that the median is not a very discriminating measure especially, for example, if only three judgment categories are used. It becomes more valuable with an increasingly greater number of categories.

A fourth method of treating judgment data consists of computing the mean stimulus value to which a given judgment is applied under different conditions, e.g., under the usual conditions (unanchored) and under anchoring conditions. Shift can then be evaluated by computing the relevant statistics. For example, the unanchored and anchored conditions may be compared by testing for the significance of difference between the means of the stimuli for the two conditions. As long as the stimuli being measured are physical in nature, continuous, and the application of numbers to the events (stimulus values) can be specified in terms of physical operations equivalent to the statistical manipulations, the problems of the first method mentioned can be successfully avoided. This method was employed by, among others, Tresselt and Becker (**11**) for the judgment of length of lines. The first disadvantage of this method appears in the way these investigators utilized it. They compared the mean length of line characterized as medium under two different conditions, leaving out the same data for the lines characterized as long or short. This was done because the response medium was the most frequent one, with fewer long and short responses. In the usual absolute judgment situation where quite often as many as nine (see Helson [**4**]) judgments are used by *S* at least some of the stimulus values, equivalent to a given response, are indeterminate because the response has not been employed by *S*. This situation becomes even more extreme in the "shifted" situation where the effect is such as to cause elimination of some of the judgments used in the "preshift" condition. It becomes obvious that this method cannot be applied to all judgments because all are not always used, and even when all are used, they do not occur with equal frequency, thus resulting in different degrees of reliability for the estimates of different judgments. What Tresselt and Becker (**11**) did to get around this problem was simply to use that one response which had the largest frequency of stimulus values to estimate the judgment value. While this is a solution of a kind, it suffers from the fact that only some of the data can be used; furthermore, while the amount of data discarded for a three-point situation is not very large it increases as the number of judgment categories increases.

A fifth method applied to judgment data is that of graphing the results in terms of percentages. Wever and Zener (**12**) plotted the percentage of times different categories of judgment were used by an *S* against the stimulus values to show that the method of absolute judgment yields the same kind of results as the method of comparative judgment. Postman and Miller (**9**) presented their data in a similar and perhaps more revealing fashion. Placing the judgment categories on the abscissa, they plotted the cumulative percentage of the occurrence of each category separately for each stimulus value; thus they arrived at as many curves for each condition as there were stimuli presented. This procedure was followed for the shift and preshift conditions so that these could be compared to determine the degree of shift. Presentation of the distribution of judgments under different conditions (e.g., unanchored and anchored conditions) yields an excellent view of the phenomenon of shift. When graphing must be done for many *S*s it becomes unwieldy; if groups are to be compared some index of relationship between the curves becomes necessary, and finally since these curves are plotted in terms of percentage a great number of presentations becomes necessary for reliable curves.

A sixth technique of treating the method of single stimuli was originated by Mausner (**6**). He made a frequency distribution of the judgments of each *S* under two different conditions of judgment ("A"—*S* judging alone; "T"—two *S*s judging in the presence of each other). He then took the difference in frequency of occurrence between the two situations ("T" − "A") for each category of judgment, e.g., if the judgment category "4" occurred 0 times in situation "T" and 18 times in situation "A" the difference for that category was −18 ("T − A" score). Then he found the midpoint of the judgment categories, e.g., for an *S* who uses judgment categories 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, the midpoint would fall exactly between judgment categories 8 and 9. Algebraic sums of the T − A scores were then obtained separately for all the judgment categories above $\Sigma(T-A)_{\text{above}}$, and separately for all the judgment categories below $\Sigma(T-A)_{\text{below}}$ the midpoint. These two sums were then totaled without respect to sign to yield a shift score.

This method is based upon the following line of reasoning: The phenomenon of shift manifests itself in an increase in frequency of use of one half of a scale of judgment and a corresponding decrease in frequency of use of the other half of the scale. The sum of the two subtotals, $|\Sigma(T-A)_{\text{above}}| + |\Sigma(T-A)_{\text{below}}|$, is then assumed to reflect the amount of shift.

This method of data treatment assumes rank order of the judgment categories but in using frequencies is free from the objections raised against the rating method.

Mausner (6) derived still another score which he named the score of direction of shift. In this method, he counted the number of plus and minus signs of the T − A scores, referred to above, separately for the judgment categories above and below the midpoint. He took the difference in frequency between the positive and negative T − A scores separately above and below the midpoint, e.g., if *S* uses 12 judgment categories (there are 6 above and 6 below the midpoint) and there are 5 negative T − A scores and 1 positive T − A score below the midpoint, this will result in a difference score of 4 negative T − A scores; if there are 6 positive and no negative T − A scores above the midpoint, then we will obtain a difference score of 6 positive T − A scores. Remembering that the phenomenon of shift manifests itself in an increase in frequency of use of one half of a scale of judgment and a corresponding decrease in frequency of use of the other half of the scale, we must add positive T − A differences above the midpoint to negative T − A differences below (negative ones above to positive ones below the midpoint). If there is a preponderance of positive T − A differences above the midpoint and/or a preponderance of negative T − A differences below the midpoint, then the direction of shift may be characterized as upward. This was true in the example given above since the 6 positive T − A differences above the midpoint must be added to the 4 negative T − A differences below the midpoint to result in a direction of shift score of +10 (where + indicates an upward shift and − indicates a downward shift).

It must be noted here that this method was designed by Mausner because his degree-of-shift score does not indicate the direction of shift. Usually such a score is not necessary. In addition, when the number of judgment categories is small the amount of discrimination possible between *S*s is small. It must be noted that since both methods rely ultimately upon a counting procedure, they have the advantage of not being open to attack from the point of view of the inequality of the distances between the categories.

The eighth and ninth techniques

of treating absolute judgment data were derived by the author for a weight-judgment technique applied to schizophrenics and normals (**10**). The first of these two techniques made use of the frequencies while the second made use of the physical scale of stimuli (weights in this case).

The frequency method consisted, first of all, of tabulating the total number of judgments according to the categories: very heavy, heavy, medium, light, and very light for the unanchored and for the anchored conditions. If the anchor makes any difference in the judgments, the frequency of certain categories should decrease and that of others should increase. A heavy anchor would tend to decrease the frequencies of the heavier judgments and increase the frequencies of the lighter judgments, and vice versa for the light anchor. A comparison of the frequency distributions by judgment categories was made between the unanchored and anchored conditions. This was carried out by using the Kolmogorow-Smirnov test (**2**). It involves a comparison of the cumulative frequencies of judgments under the two conditions. The maximum discrepancy found between the two cumulative frequency distributions for each *S* is a measure of amount of shift. This score will be designated as the *category-shift score*.

An example of the manner of calculation of the category-shift score is given below:

a. Tabulation of the frequency of judgments in each condition (unanchored, heavy anchor, light anchor) separately for each *S*, e.g., subject X,

	VL	L	M	H	VH
NA	5	7	4	6	3
HA	8	10	6	1	0

where VL = very light, L = light, M = medium, H = heavy, VH = very heavy; NA = unanchored condition, HA = heavy anchor condition.

b. Cumulative frequency distribution from VL to VH of the frequencies in each category for all conditions; following the above example the table under *a* would be transformed into the table below:

	VL	L	M	H	VH
NA	5	12	16	22	25
HA	8	18	24	25	25

where the entry in each cell now represents the frequency of judgments of the judgment category to which the cell refers plus all the frequencies of all the judgments lighter than the one under consideration.

c. Subtract 'the cumulative frequencies of the appropriate NA from the HA conditions (the LA—light anchor condition—from the appropriate NA conditions) and use the largest difference as the shift score. Following the above example:

	VL	L	M	H	VH
HA	8	18	24	25	25
NA	5	12	16	22	25
D	3	6	(8)	3	0

where D = difference between the HA and NA conditions and the number in parentheses represents the maximum difference between the two cumulative frequency distributions. This difference is the category-shift score. If an investigator so desires, he can evaluate the statistical significance of the shift separately for each *S*. Goodman (**2**) provides a table for this purpose. If interested in comparing groups one can use the maximum difference between cumulative distributions (the category-shift score) as a score for each *S*. These scores which are frequencies can then be manipulated statistically. This technique like all methods making use of frequency can be manipulated statistically without fear of dealing with scales of unequal intervals as in the rating scale method. It has the advantages of being simple in calculation and of providing the investigator with an immediate estimate of statistical significance.

The ninth method of computing the shift score for each *S* used the difference between the stimulus (weight) to which a particular judgment was assigned and the one to which it should have been assigned (i.e., the correct weight) according to prior verbal instructions given to *S*. These differences were then summed separately for judgments assigned to weights heavier and weights lighter than the one to which they should have been assigned. The difference between the two differences resulted in a separate score for the anchored and unanchored conditions; the difference between the anchored and unanchored condition scores in turn gave rise to the shift score which will be designated as the *stimulus-shift score.*

An example of the manner of calculation of the stimulus-shift score is given below:

a. The responses (judgments of) to each of the weights were tabulated as shown below, separately for each NA, HA, and LA condition, e.g., subject X;

			NA		
	VL	L	M	H	VH
200	4	1			
250	1	3	1		
300		3	2		
350				4	1
400			1	2	2

			HA		
	VL	L	M	H	VH
200	4	1			
250	4	1			
300		5			
350		3	2		
400			4	1	

b. Inspection of the tally tables for the unanchored condition (NA) and heavy anchor condition (HA) shows all the "correct" judgments along the diagonal, that is, all the responses that have been attributed to the stimuli to which *S*s were instructed to attribute them. Above the diagonals are responses that have been attributed to stimuli *lighter* than the ones to which *S*s were previously verbally instructed to attribute them, while below the diagonal are the responses that have been attributed to stimuli *heavier* than the ones to which *S*s were previously instructed to attribute them. Thus it was possible to obtain two scores for each condition, namely the number of responses attributed to stimuli heavier and the number attributed to stimuli lighter than the stimuli to which they should have been attributed according to previous instructions. To get a more exact measure of discrepancy between judged and actual weight, the difference in grams between the actual and the judged weight was obtained.

For example, looking at the entry in the NA table above, the entry in Cell 200-L gave rise to a discrepancy score of 50 gms., since the judgment light was attributed to a weight 50 gms. lighter than the one to which it should have been attributed according to previous instructions; the same holds for the entries in the Cells 250-M and 350-VH. By adding these three discrepancy scores, a total discrepancy score "lighter" of 150 gms. is obtained.

Below the diagonal, discrepancy scores can be obtained in an analogous manner. Cell 250-VL gives rise to a discrepancy score of 50 gms. because the response VL was attributed to a weight 50 gms. heavier than the one to which it should have been attributed according to previous instructions. In Cell 300-L there are three responses that have been attributed to the same stimulus that is 50 gms. heavier than the one to which they should have been attributed according to previous instructions, thus giving rise to a discrepancy score of 150 gms. The Cell 400-M shows a discrepancy of 100 gms. because the response M was attributed to a weight 100 gms. heavier than the stimulus to which it should have been attributed according to previous instructions. Finally, Cell 400-H shows a discrepancy score of 100 gms. because there are two responses that have been attributed to the same stimulus that is 50 gms. heavier than the one to which they should have been attributed according to previous instructions. By adding the above four discrepancy scores a total discrepancy score "heavier" of 400 gms. is obtained.

By means of the same procedure two total discrepancy scores can be obtained from the HA table. The discrepancy score "lighter" is 50 gms. while the discrepancy score "heavier" is 1300 gms.

c. Subtract the total discrepancy score "lighter" from the total discrepancy score "heavier" for each condition, thus obtaining an estimate of the bias or net tendency to make errors in the direction of attributing judgments to weights heavier or lighter than the ones to which they should be attributed according to instructions. In the example given, the net bias score for condition NA is 400 gms. −150 gms. = 250 gms., while the net bias score for condition HA is 1300 gms. − 50 gms. = 1250 gms.

d. Finally to obtain a shift score subtract the net score of NA from that of HA (of LA from NA). In this case the subtraction of 250 gms. from 1250 gms. yields a stimulus shift score of 1000 gms. for subject X.

This last score made use of the physical scale underlying the judgment scale. It is based on obtaining the difference in the stimulus dimension under consideration, that is, between the stimulus being judged at any given time and the stimulus to which the judgment correctly belongs. This shift score takes into account not only the direction of shift and the frequency with which various categories are used but also considers the degree of shift, i.e., how many gms. difference there is between the stimulus being judged and the stimulus which *S* thinks he is judging.[3]

Since both the category-shift score and the stimulus-shift score were computed on the same set of data, it was possible to compare them. Table 1 provides us with the rank-order correlations between the two types of shift scores, computed on 16 normals and 16 schizophrenics during different experimental sessions and due to different anchors. Since the coefficients are high either score can be substituted for the other. Further inspection of the scores makes plain, however, that the stimulus-shift scores show greater discrimination than the category-shift scores. Amount of discrimination between *S*s can be roughly measured in terms of the number of tied scores for *S*s. The total of such tied scores for the weight-shift score over all *S*s and

[3] A measure of kinesthetic sense can be established by adding the total discrepancy scores "lighter" (above the diagonal) and "heavier" (below the diagonal) obtained in computing the shift measure; the smaller this total is, the better the kinesthetic sense. Thus, this score is really an error score. Using the example given for calculation of the shift measure, a kinesthetic score of 150 gms. +400 gms. = 550 gms. is obtained for the NA condition and a score of 50 gms. +1300 gms. = 1350 gms. is obtained for the HA condition.

conditions was 49 while that for the category-shift score was 107.

In conclusion, it can be stated here that while some of the criticisms (like those made against the direct use of ratings in statistical manipulations based on assumptions of equal intervals) would advise against any use of the method of computing a shift score, most of the criticisms are of the nature of specifying under what conditions a particular method might or might not show itself up to advantage.

TABLE 1

Rank-Order Correlation Coefficients Between Two Methods of Measuring Shift for 16 Equated Normals and Patients as a Function of Heavy and Light Anchor Conditions in Two Successive Sessions

Session	Condition	Normals	Patients
1	Heavy anchor	.91**	.95**
	Light anchor	.96**	.93**
2	Heavy anchor	.92**	.83**
	Light anchor	.95**	.79**

** Significant beyond the .01 level.

REFERENCES

1. Brown, D. R. Stimulus-similarity and the anchoring of subjective scales. *Amer. J. Psychol.*, 1953, **66**, 199–214.
2. Goodman, L. A. Kolmogorow-Smirnov Tests for psychological research. *Psychol. Bull.*, 1954, **51**, 160–168.
3. Heintz, R. K. The effect of remote anchoring points upon the judgment of lifted weights. *J. exp. Psychol.*, 1950, **40**, 584–591.
4. Helson, H. Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychol. Rev.*, 1948, **55**, 297–313.
5. Likert, R. A. Technique for the measurement of attitudes. *Arch. Psychol.*, 1932, No. 140.
6. Mausner, B. The effect of prior reinforcement on the interaction of observer pairs. *J. abnorm. soc. Psychol.*, 1954, **49**, 65–68.
7. McGarvey, Hulda R. Anchoring effects in the absolute judgment of verbal materials. *Arch. Psychol., N.Y.*, 1943, No. 281.
8. Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, **1950.**
9. Postman L., & Miller, G. A. Anchoring of temporal judgments. *Amer. J. Psychol.*, 1945, **58**, 43–53.
10. Salzinger, K. Shift in judgment of weights as a function of anchoring stimuli and instructions in early schizophrenics and normals. Unpublished doctor's dissertation, Faculty of Pure Science, Columbia Univer., 1954.
11. Tresselt, Margaret E., & Becker, M. Scales of judgment and personality correlates. *J. gen. Psychol.*, 1950, **43**, 221–230.
12. Wever, E. G. & Zener, K. E. The method of absolute judgment in psychophysics. *Psychol. Rev.*, 1928, **35**, 466–493.

Received May 4, 1956.

Psychological Bulletin
Vol. 53, No. 5, 1956

MASS SCREENING AND RELIABLE INDIVIDUAL MEASUREMENT IN THE EXPERIMENTAL BEHAVIOR GENETICS OF LOWER ORGANISMS

JERRY HIRSCH[1]
Columbia University

AND ROBERT C. TRYON
University of California

At our present stage of ignorance about how genes determine behavior, we might well concentrate on experimental studies of lower organisms. Their reactions may be thought of as the emergent behavior which has developed through evolution into the complex behaviors of higher organisms. Knowledge gained from such studies may provide conceptual models leading to an understanding of how hereditary and stimulus components interact in determining higher forms of behavior.

For this purpose the use of lower organisms offers distinct advantages. There is a brief time span between generations, permitting *E* to perform in a short time period the various crossings essential to fundamental genetic studies. Each generation produces abundant progeny, enabling *E* to recover the extreme behavior types required in selective breeding experiments. And further, the genetics of their morphology is better understood than is that of higher forms. The fruit fly, *Drosophila*, has all of these advantages.

First, however, reliable techniques for measuring individual differences (hereafter referred to as *ID*s) in behavior must be developed. Reliability coefficients *must* be calculated, and they must be *high*. The problem reduces to the question: How can we observe the behavior of large numbers of very small *S*s and at the same time reliably measure the performance of each *S*?

This paper presents a method which accomplishes both these objectives. We call it the method of "*mass screening with reliable individual measurement*." As an illustration of the method, we will show that in the mass observation of a particular behavior of *Drosophila*, reliability coefficients of about .9 can be secured in an experimental test period of four minutes. During this time 15 sample observations of 15 sec. each were made. Each individual was observed as a member of a group of other flies. The method shows that *Drosophila ID*s can be measured as reliably as human *ID*s. Indeed, we know of no experiment on men covering 15 brief observations that yields a reliability as high as .9.

Genetics has up to the present concerned itself with physical characteristics rather than with behavior. The reliability of individual measurement is not so obviously important in the study of morphological characteristics; usually the characteristic is either present or absent, or present in only a small number of forms, and its presence or absence is immediately obvious, (e.g., eye color, notched wing, bar eyes, etc.). Individual differences in behavior, on the other hand, are not so easily recognized:

[1] National Science Foundation Postdoctoral Fellow at the University of California, Berkeley.

such recognition requires special methods.

There are at least three reasons why we need reliable measurement of such *ID*s:

1. Reliable phenotypic differentiation is needed for selective breeding for homozygous lines. Both the purity of different strains and the rapidity of selection are limited by our capacity to discriminate between individuals, since, as the errors of measurement decrease, the probability increases that individuals with the same score will be genetically similar.

2. The study of learning also requires reliable individual measurement because of the relation between the strength of the unconditioned response and conditioning.[3] (Obviously for those individuals in whom the unconditioned response has zero strength, conditioning is impossible.) We believe that the study of learning requires reliable knowledge of the distribution of *ID*s in the population being sampled. Much effort has been spent in demonstrating the influence of environment on behavior. It is patent, however, that environmental influence must be an influence on something and therefore the laws of such influence must differ as the object influenced differs.

3. Reliable individual measurement is essential for answering three questions about the generality of any behavior: (*a*) Temporal generality; how long does a given disposition to respond endure and to what extent does the rank ordering of individuals persist over this period? (*b*) Stimulus generalization; over what range of stimuli can the response be evoked and how well is the rank ordering of individuals maintained over that range? (*c*) Behavior generality; to what extent do other behaviors preserve the rank ordering of individuals?

Efficient methods of observation are also a desideratum for studying small organisms. It is a theorem in sampling theory that the detection of extreme cases, a necessity in genetic selection experiments, requires the observation of large numbers of *S*s since the probability of finding these extreme cases is a direct function of the sample size. Rapid observation permits the examination of large numbers of *S*s and thus increases the sampling stability essential to the generality of the findings. Furthermore, replication of experiments can be undertaken without excessive labor.

The next section of this paper presents a method for reliably measuring *ID*s in behavior by means of *mass screening*, a procedure which achieves the objective of reliably *classifying every individual's behavior without handling or observing each small organism individually*. The method is completely general and easily applicable to the study of any behavior, both unconditioned and conditioned.

This objective is illustrated by the results of an experiment that employed the mass screening technique in the study of the geotropic reactions of *Drosophila melanogaster*. A series of 15 successive mass screenings, for example, produced 16 test tubes, each containing a different geotropic class of *Drosophila*. The flies in the tubes 0 to 15 represent different degrees of the negative geotropism. That is, the flies are differentiated on this final composite 16-point scale based on 15 prior mass screenings in which the individuals were not separately han-

[3] Use is made of conditioned response terminology for convenience of exposition. It is not intended to represent a theoretical statement about the nature of the learning process.

dled. The reliability coefficient of this final scale score is determinable and in principle, it can be increased to any desired value by further mass screenings.

Experimental Design and Analytic Procedures

The method consists of cumulating a total composite score X_t, for each organism in any behavior, X, where:

$$X_t = X_1 + X_2 + X_3 + \cdots + X_n.$$

$X_1, X_2, \cdots, X_n$ represent scores earned by it in n comparable sample mass screenings. Setting up such a total score is the essence of psychological test theory. Most of the formulae used in this paper are standard in psychological test theory. A simple summary of them can be found in J. P. Guilford's *Psychometric Methods*, Chaps 13, 14, 15 (**1**). Guilford's rationale of the formulae, however, is based on the factorial truth-error doctrine. In another paper one of the authors develops them with fewer assumptions (**3**). Our procedure adapts these principles to the problem of calculating reliability coefficients for the scores of individuals who are only observed as members of a large group.

The main steps of the procedure are as follows:

1. Conceptualize the behavior property, X, that is to be scaled, and operationally define it with sufficient specification to indicate the general conditions under which it may be observed.

2. Devise a standard test sample procedure for obtaining a unit measure of *ID*s in X, one which has the advantage of permitting observation of a large group of *S*s at one time while locating the total, N, of individuals in subgroup classes scored 0, 1, 2, $\cdots$, k in magnitudes of X.

3. Take a randomly bred sample of the *S*s and mass screen them through n replications of the standard procedure. At the end of every replication, score each subgroup by its cumulative total score, X_t, then combine subgroups with the same X_t score and proceed with the next replication.

4. Calculate the reliability coefficient, r_{tt}, of each successive X_t score, decide on the value of n which will yield a reliability of sufficiently high magnitude, then examine the shape of the distribution of the X_t scores of the individuals.

5. If the original method results in a low reliability or an excessively skewed distribution of final composite scores, alter the standard test, take a second random sample and repeat the general procedure. Several such experiments may be required before an adequate method of observation is discovered.

The details of the steps of this *general* procedure will be developed and illustrated by an experiment conducted by one of the authors on *ID*s in the geotropic reaction of *Drosophila.*

1. *Conceptualization and Definition of the Behavior*

The behavior chosen was the unconditioned disposition to go in the direction opposite to gravity. This negative geotropism is operationally defined as an upward movement of the fly whenever it is placed in any situation permitting travel upward, other external stimuli which might induce vertical movement being controlled.

2. *Standard Test Sample Procedure*

The test situation consists of two test tubes, a lower one standing upright in a rack, the other inverted over the mouth of the lower one. Since the flies are also phototropic, the light source was placed at right angles to the vertical. A group of flies are placed in the lower tube, shaken to the bottom, and then allowed to ascend. At the end of an arbitrary "cutting point" time of 15 sec., a card is inserted between the lower and the upper tubes. The upper tube is scored and labeled "1," and lower tube "0."

Thus the standard sample observation in this case is like a dichotomous test item, the top tube scored "pass" and the lower one "fail." A cutoff point of 15 sec. was found empirically to divide the group of flies into two approximately equal pass and fail subgroups, a division which avoids skewness in the distribution of final composite X_t scores.

It should be emphasized that dichotomous scoring is ***not*** a necessary restriction of the method. The standard procedure could have been devised to provide more classes. The pass-fail break was chosen for experimental convenience.

This standard test procedure, though satisfying the operational definition of geotropism, might not elicit uniquely a systematic reaction to gravity. Since the test tube situation permits only movement upward it may be that, if there is an *activity* differential among the *S*s, the flies that are upwardly mobile may be very active flies. Only additional experiments which control activity can resolve the matter. Thus, we use the term "geotropism" here only in an operational sense, recognizing that the *ID*s observed in this situation might later be shown to be significantly influenced by additional components.

3. *Choice of an Unselected Sample*

Since the range and reliability of *ID*s is partly a function of the heterogeneity of the *S*s, a stock of unselected *Drosophila* with a history of random mating was chosen.

4. *Mass Screening*

A random sample of 106 flies was screened and scored by the following procedure.

a. First composite score, $X_{t_1}=X_1$. The results of the first observation are shown in Fig. 1, which reproduces part of the score sheet actually used. Under X_1 and f_1 it can be seen that 54 flies ascended to the upper tube, earned a "pass" and thus received a score of $X_1=1$. There are 52 flies that remained in the lower tube, earned a "fail" and received a score of $X_1=0$. The scores, X_{t_1}, of this trial take the values of 1 and 0.

b. Second composite score, $X_{t_2}=X_{t_1}+X_2$. The 54 flies with $X_{t_1}=1$ were put through the standard procedure a second time for Trial 2. The 46 flies that ascended earn a tube score, $X_2=1$, and a composite score $X_{t_2}=2$; the 8 remaining down have $X_2=0$ and $X_{t_2}=1$, as shown. In similar fashion the flies with $X_{t_1}=0$ divide into 22 earning $X_2=1$, $X_{t_2}=1$ and 30 earning $X_2=0$, $X_{t_2}=0$.

c. Third composite score, $X_{t_3}=X_{t_2}+X_3$. The standard procedure is repeated for each of the three X_{t_2} classes resulting from Trial 2.

Note, even though there are four X_2 tubes of flies at the end of Trial 2, there are only three X_{t_2} classes. The two subgroups with 8 and 22 flies have been combined in one tube because both received the same score, $X_{t_2}=1$, i.e., the same composite score is the cumulative sum of all previous scores irrespective of the order in which the individual "passes" and "fails" were obtained.

d. Additional composite scores, X_{t_4}, X_{t_5}, · · · . The procedure is continued by taking further sample observations; at the end of each one, subgroups having the same X_t score are

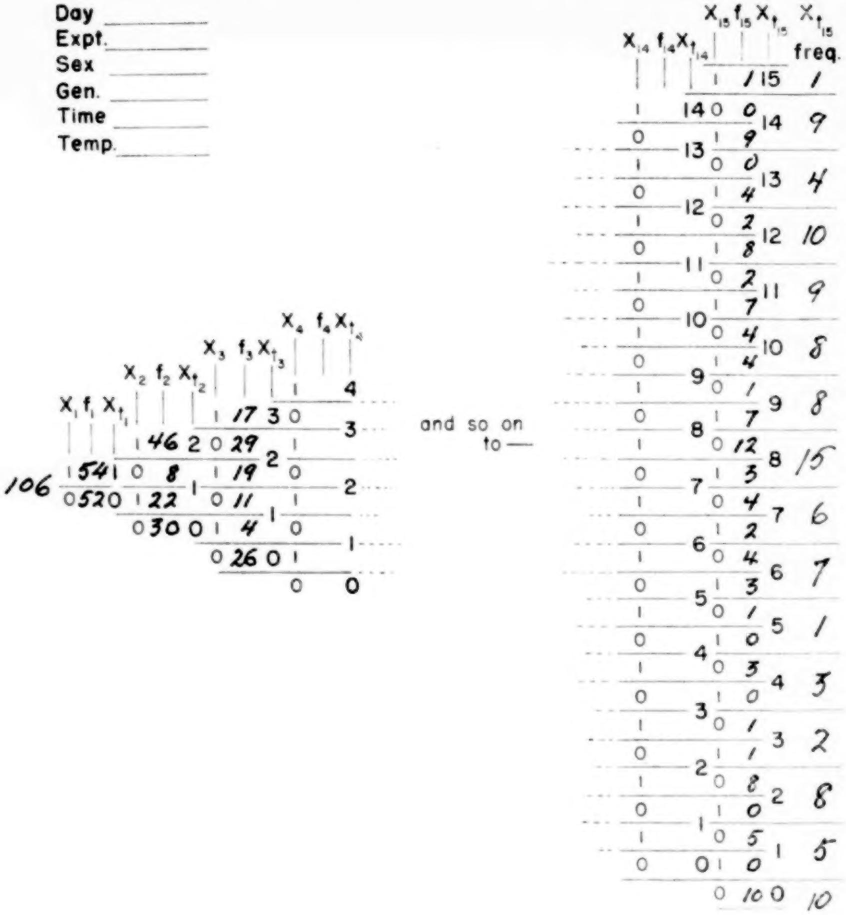

FIG. 1. MASS SCREENING SCORE FORM

combined for the next observation. Figure 1 shows the results schematically up through $X_{t_{15}}$.

The reason for the "experimental convenience" of dichotomous classes in the standard procedure should now be apparent; with more than two classes the number of subgroups becomes unmanageable.

5. *Analysis*

a. The distribution of X_t scores. One of the objectives of experimental behavior genetics is reliable differentiation between individuals and subsequent genetic validation of differences by means of selective breeding. Since, for a given behavior, it is assumed that there is a range of ability and that the *S*s in a population are distributed over the range, it follows that any methods which tend to pile up the final scores in a few extreme categories should be eschewed in favor of others which distribute the scores more widely. The individuals whose behavior is under observation will be used for breeding, hence it is important to differentiate them clearly on the behavioral scale. Failure to do this prevents the discovery of any genotypic differences that might exist.

The *E* can usually control the form of the distribution of total X_t scores. In our illustrative experiment this control was accomplished through selection of the time interval in which the response can be performed, i.e., the proportions p, of "passes" and q, of "fails" vary as a function of the amount of time allowed in the test tube. In examples from several experiments it may be shown that when $p > .5$, the X_t distribution is negatively skewed and when $p < .5$, X_t is positively skewed. Either type of skewness is undesirable because cases pile up in the extreme categories where, for the purposes of selective breeding, the finest differentiations are needed.

This point is illustrated in Table 1 where the frequency distribution of the composite score $X_{t_{10}}$ from Fig. 1 is presented in the first row of entries. A 15-sec. cutoff was used for this sample. The mean proportion earning a score of $X_1 = 1$ on the ten successive standard tests is $\bar{p} = .5$. The distribution is seen to be platykurtic with no appreciable piling up of the cases in the extreme categories. This is the result of the approximately 50–50 cut on each trial.

The effects of extreme cuts are shown in the other rows of Table 1. For the group with an 8-sec. cutoff in the standard test the proportion getting into the upper tube is $\bar{p} = .16$, with the result that the composite $X_{t_{10}}$ scores are very positively skewed with a pile up of flies in the 0 category. The opposite extreme cut of 27 sec. gives a $\bar{p} = .66$, with a pile up at the high $X_{t_{10}}$ scores.

b. Reliability of X_t scores. It is important that the composite X_t score be reliable if *E* is to use the differentiations between individuals as the basis for further experimental work on selective breeding, conditioning, or the investigation of the generality of behavior *X*. The reliability coefficient, r_{tt}, cannot be computed by the split-half method in the mass screening method because combining into a single group all *S*s with the same composite X_t score loses the specific sample score history of each individual. The coefficient can be estimated accurately, however, from the variances of the composite X_t score and of the individual test sample scores, as follows (**3**, Formula 12):

$$r_{tt} = \frac{n}{n-1}\left(1 - \frac{\Sigma V_i}{V_t}\right), \qquad [1]$$

where:

n = number of standard test samples or replications.

$\sum V_i$ = sum of the variances (σ_i^2) of the n test samples.

TABLE 1

DISTRIBUTION OF INDIVIDUALS IN COMPOSITE X_t SCORE

(Entries are frequencies)

p	Cutoff	$X_{t_{10}}$											
		0	1	2	3	4	5	6	7	8	9	10	N
.50	15 sec.	11	5	9	7	10	8	11	13	15	11	6	106
.16	8 sec.	54	13	8	8	7	5	4	2	3	0	0	104
.66	27 sec.	0	9	3	3	4	4	4	16	24	12	13	92

V_t = variance of the final composite X_t scores, i.e., σ_t^2.

When, as in the present case, the standard procedure gives a dichotomous cut, the variance, V_i, of any particular sample observation is:

$$V_i = pq, \qquad [2]$$

where:

p = proportion of individuals above the cut in all subgroups
= mean score when, as in the example, those above the cut are scored 1, those below 0.
$q = 1 - p$.

The values of the reliability coefficients and of other constants for several *Drosophila* experiments are given in the third rows of Table 2. The first group is the one presented in Fig. 1, in which 15 sample observations were finally taken under conditions believed to produce optimum differentiation between individuals. It will be noted that, beginning with the fourth column of entries, after the first few "adjustment" trials the reliabilities progressively increased to .87 for the final composite score based on 15 sample observations.

The *E* naturally asks: are the successive sample observations strictly comparable measures of the property *X*, here the negative geotropic reaction? The additional constants of Table 2 give insight into this question.

If the individuals systematically improve or deteriorate in performance the mean score, p_i, and the variance, $V_i = pq$, of successive observations will both change. In the first and second rows of Table 2 we see that in our example p_i and therefore V_i both remain relatively constant.

If the individuals become either more reliably differentiated or less so as screening proceeds, then the reliability coefficient will not increase according to the "Spearman-Brown law" of increased reliability with the addition of comparable sample observations. Evidence on this point can be secured in two ways.

TABLE 2

RELIABILITY COEFFICIENTS AND OTHER CONSTANTS IN THE DROSOPHILA GEOTROPIC EXPERIMENTS
SAMPLE OBSERVATION, X_i

Group	n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
15 sec. $N=106$	p_i	.64	.38	.50	.53	.47	.53	.55	.63	.55	.48	.49	.48	.46	.46
	V_i	.23	.24	.25	.25	.25	.25	.25	.23	.25	.25	.25	.25	.25	.25
	r_{tt}	.62	.49	.60	.70	.75	.78	.80	.82	.82	.83	.83	.84	.86	.87
	n_c	23	60	49	42	39	39	39	39	42	42	44	47	44	44
	$\bar{r}_{ij}$	.45	.24	.28	.31	.33	.33	.33	.33	.31	.31	.30	.29	.30	.30
8 sec. $N=104$	p_i	.15	.20	.11	.18	.16	.18	.17	.18	.21					
	V_i	.13	.16	.10	.15	.14	.15	.14	.15	.17					
	r_{tt}	.66	.63	.68	.68	.72	.75	.77	.80	.81					
	n_c	20	34	35	47	44	47	44	44	44					
	$\bar{r}_{ij}$	.49	.36	.35	.29	.30	.29	.30	.30	.30					
27 sec. $N=92$	p_i	.83	.67	.71	.71	.67	.71	.62	.57	.57					
	V_i	.14	.22	.21	.21	.22	.21	.24	.25	.25					
	r_{tt}	−.04	.15	.51	.59	.69	.72	.72	.76	.78					
	n_c		298	76	67	44	44	57	57	54					
	$\bar{r}_{ij}$	−.02	.06	.20	.22	.27	.27	.25	.25	.26					

The first is to discover whether the mean correlation, $\bar{r}_{ij}$, between sample observations entering into the composite, X_t, changes for successive X_t scores. From the familiar Spearman-Brown approximation (**3**, Formula 17), we note that the reliability coefficient, $r_{t_nt_n}$, for any composite X_t based on n samples is:

$$r_{t_nt_n}=\frac{n\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}}, \qquad [3]$$

whence, solving for $\bar{r}_{ij}$:

$$\bar{r}_{ij}=\frac{r_{t_nt_n}}{n-(n-1)r_{t_nt_n}} \qquad [4]$$

The successive values of $\bar{r}_{ij}$ are given in Table 2, fifth rows. We note that after the first few trials $\bar{r}_{ij}$ plateaus around .30.

The other way is for E to set a desired reliability for the final composite, and solve for the value of n in Equation 3 that will achieve this desired reliability. Suppose E desires a reliability of .95. Call this R_{tt}. Set R_{tt} into Equation 3 and solve for n:

$$n=\frac{R_{tt}(1-\bar{r}_{ij})}{\bar{r}_{ij}(1-R_{tt})}. \qquad [5]$$

The values of n for $R_{tt}=.95$ are given in the fourth rows of Table 2. In general they remain around 45 trials. This finding has the practical value of informing E how many sample trials are necessary to achieve the reliability he desires. If n turns out to be too large, a design having more classes per trial might be considered as a means of reducing the number of trials required.

When the individual test sample scores are not available, as is the case when groups are screened on the multiple-unit discrimination maze (**2**), the reliability coefficient can be computed directly from the final distribution of X_t scores by means of the Total Score formula (**3**, Formula 37):

$$r_{tt}=\frac{n}{n-1}\left(1-\frac{M_t-M_t^2/n}{V_t}\right), \qquad [6]$$

where M_t equals the mean of the final composite X_t scores.

c. Domain validity coefficient of the composite score, X_t. The reliability coefficient, r_{tt}, though necessary in the above formulations, is not the best statement of the reliability of the composite X_t. A more meaningful index is the correlation between the X_t scores and that on an indefinitely large number of screenings, namely X_{t_∞}. Though the "true score," X_{t_∞} is not available, the correlation r_{tt_∞} can nevertheless be estimated as follows (**3**, Formula 21):

$$r_{tt_\infty}=\sqrt{r_{tt}}. \qquad [7]$$

Thus, in our case our X_t based on fifteen screenings would correlate $r_{tt_\infty}=\sqrt{.867}=.93$ with a perfectly reliable measure based on many such screenings. This coefficient also has the following added meaning: If we had the true score of each fly based on many sets of 15 screenings, the ratio of the standard deviation of these true scores to that of the observed X_t score would be .93. In short, the distribution of true scores would look much like that actually observed.

d. Individual variance ("errors of measurement"). In order to conduct experiments on selective breeding, conditioning, or generality it is necessary to get a practical estimate of the amount of difference in X_t scores among individuals that is undetermined, i.e., not assignable to known sources of variation. This estimate is the individual variance, V_o (**3**, Formula 23a), where:

$$V_0 = V_t(1 - r_{tt}). \qquad [8]$$

In our example for $X_{t_{15}}$, $V_t = 4.40$, hence the individual standard deviation is:

$$\sigma_0 = 4.40\sqrt{1 - .867} = 1.6.$$

The necessity of a high reliability can be seen in the above formula: as the reliability approaches unity the amount of variation attributable to individual variance tends to vanish.

Nonuniformity of individual variance. The individual variation, however, is most likely not constant over the final distribution: (*a*), an extreme score can vary in only one direction, towards the mean: (*b*), the individuals receiving extreme scores have shown perfectly consistent performance throughout, that is, either they have always scored a zero or they have always scored one. Hence, it might be expected that the individual variation, as estimated by a retest, should be much smaller at the extremes than in the middle of the distribution.

Empirical check. To assess this possibility a retest or validation experiment may be performed. In our illustration, the *S*s receiving extreme X_t scores of 15 and 14 were combined and put through $n' = 10$ additional trials; also those receiving middle X_t scores of 7 and 8 were put through a retest of 10 trials. For the extreme categories $\sigma_{t_e'}^2 = 4.00$, while for the middle categories $\sigma_{t_m'}^2 = 5.82$, the latter being significantly larger than the predicted variance for the middle categories. It is evident that the assumption of uniformity of individual variance over the whole X_t scale is doubtful.

Limits of Selective Breeding

How many generations is it necessary or practical to continue a selective breeding program, i.e., what are the criteria for stopping? The individual standard deviation, $\sigma_o = \sqrt{V_o}$, provides an answer to this question: it is useless to attempt further selection in any line beyond the point where its $\sigma_t = \sigma_o$; at that point the method of observation no longer reliably differentiates individuals, i.e., neither selection nor the evaluation of the results of selection are any longer possible. In our case, no further selective breeding would be attempted in any line whose σ_t was much below 1.6.

Summary

Fast breeding, prolific, small organisms are pre-eminently suited for studies in the field of behavior genetics. Their value as experimental *S*s is further enhanced by the method of mass screening that succeeds in combining the objective of reliable individual measurement with that of mass observation. Hence, it is now possible to achieve the experimental desiderata of efficiency, reliability, and brevity in the field of behavior genetics. The method is illustrated by experiments on the geotropic responses of *Drosophila*.

REFERENCES

1. Guilford, J. P. *Psychometric methods.* (2nd Ed.) New York: McGraw-Hill, 1954.
2. Hirsch, J. A multiple-unit discrimination maze for the reliable mass screening of small organisms. *J. comp. physiol. Psychol.*, in press.
3. Tryon, R. C. Reliability and domain validity: reformulation and historical critique. *Psychol. Bull.*, in press.

Received May 8, 1956.

GEORGE BANTA COMPANY, INC., MENASHA, WISCONSIN